International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 18, Number 1

February – 2017

A Computational Method for Enabling Teaching-Learning Process in Huge Online Courses and Communities

Higinio Mora, Antonio Ferrández, David Gil, and Jesús Peral
University of Alicante, Spain

Abstract

Massive Open Online Courses and e-learning represent the future of the teaching-learning processes through the development of Information and Communication Technologies. They are the response to the new education needs of society. However, this future also presents many challenges such as the processing of online forums when a huge number of messages are generated. These forums provide an excellent platform for learning and connecting students of the subject, but the difficulties in following and searching the vast volume of information that they generate may produce the opposite effect. In this paper, we propose a computational method for enabling the educational process in huge online learning communities. This method analyses the forum information through Natural Language Processing techniques and extract the main topics discussed. The results generated improves the management of the forums, increases the effectiveness of the teachers' explanations and reduces the time spent by students to follow the course. The proposal has been complemented with a real case study that shows promising results.

Keywords: natural language processing, topic detection, distance learning, online forum, mooc managing, e-learning, applications for open and distance learning

A Computational Method for Enabling Teaching-Learning Process in Huge Online Courses and Communities

The teaching-learning processes have been renovated in recent years, mainly due to technological changes and the new educational needs of society. The development of Information and Communication Technologies (ICTs) has offered new ways of interaction among students and teachers. A key element in this progress has been the platforms for Open and Distance Learning (ODL) (United Nations Educational, Scientific, and Cultural Organization {UNESCO}, 2002). These sites provide educational infrastructures to a large number of learners. In this way, the cost structure of teaching is evolving from the substitution of tangible and costly infrastructures (i.e., buildings) in the form of ICT infrastructures and intangible assets such as knowledge.

The ODL platforms are usually implemented as online web platforms for education. Learning platforms that support the electronic learning must have all the facilities for integrating the services used in the educational environment, therefore, web platforms that support the ODLs have come from being a static tool for one-way transmission of content to allow interaction and collaboration of its participants. The resurgence of the web as a collaborative platform where users can share information and take advantage of the interactions of other users to enhance their experience provides numerous benefits such as increase of the student's motivation and creation of collective intelligence (Mora, Signes Pont, & DeMiguel Casado, 2014; Masud, 2015; Lucas de Azevedo & Borges, 2015).

This continuous development of web possibilities for the creation, communication, and sharing of contents enhances the teaching-learning process and increases effectivity of learning systems for the knowledge society (Lytras, Mathkour, Abdalla, Wadee, Yanez-Marquez, & Siquera, 2015; Sicilia & Lytras, 2005). However, there are still challenges that should be specifically addressed in order to achieve the effective ODL provision and to maintain excellent educational standards. One of the most powerful and popular tools are online forums (Mora, Signes Pont, DeMiguel Casado, & Gilart Iglesias, 2015). Online forums are a communication technology tool in education and provide an excellent platform for learning and connecting students to the subject. The expected roles of the forums are to increase engagement of students in the subject, promote deep learning, and maintain motivation (Onah et al., 2014).

Technological advances and computing methods may provide tools for enhancing these roles and facilitating the educational process. In this way, artificial intelligence approaches based on Natural Language Processing (NLP) should deliver tools specifically aimed at students and teachers to make the assimilation of content and course tracking easier for students, and to provide mechanisms to manage the learning platforms by teachers. The working hypothesis of our research is that this kind of solution can be used to better engage the students and teachers in online platforms with huge forums by enabling the knowledge management and learning-teaching processes. Under our approach, the main objective of our work is to provide automatic tools and applications to facilitate tracking and monitoring online learning communities through the automatic processing of forums of lexical, syntactic, and semantic analysis; clustering; and information retrieval techniques.

The remainder of the paper is structured as follows: Section 2 describes the problem of management and monitoring forums in huge online courses and communities; Section 3 summarizes the most relevant related work on forum analysis through NLP techniques; Section 4 introduces our proposal for automatic processing in order to address the challenge and facilitate the tracking and monitoring of forums; Section 5 illustrates the application of the method on a case study. Finally, the main contributions and our directions for future works are explained in Section 6.

Online Forum Management Issue

This section introduces the main challenge of forum management in ODL platforms on which is focused the research of this work.

The ODL platforms have become very popular over the past few years thanks to the wide coverage of new technologies and the greater ability of people using the Internet. The absence of geographical and temporal barriers has caused two complementary effects: a wide range of online courses and a large number of students demanding distance learning.

Forums are a central tool in many online educational platforms. However, the advantages that this tool provides very often are attenuated due to several problems which are very demanding and time consuming. It should be noted that the forums with a great number of users and great volume of posts generated by them, cannot be monitored completely. Both students and teachers may not have time to spend on reading all posts generated each week by users in the forums. This problem has been underlined in many related works. This occurs especially in online platforms directed towards large population groups such as generic forums (i.e., StackOverflow, Quora) or Massive Open Online Courses (MOOCs) (Ramesh, Goldwasser, Huang, Daumé, & Getoor, 2014). The review presented by Hew and Cheung (2014) recapitulates the state-of-the-art on the use of MOOCs by both instructors and students in order to find the main motivations and challenges of using MOOCs. In addition, they detect a lack of student participation in online forums and identify issues that have yet to be entirely addressed or fixed. The work presented in Renz et al. (2014) introduces the major problems to face thousands of e-assessment submissions.

The huge number of messages that can be generated may cause difficulties with the online forums. At the most basic level, forums are not good at managing high volumes of posts (Lentell & O'Rourke, 2004). This situation causes the topics to become fragmented over many threads and posts. The traditional approaches are not available any longer in the current situation of online education and it needs to invest huge resources. The administration of forums leads us to wonder about how to find ways of automatic or semi-automatic management of threads in online course forums.

At the beginning of the course, teaching staff can organize the forum in several categories. In fact, this is the normal practice. But, in numerous online courses, the problem remains unsolved because so many posts can be sent in each category. In addition, this initial structure can be valid for the main themes of the subject, but other interesting topics can arise throughout the course which don't fit in any category, and posts about the same topic can be distributed in many categories.

These situations cause the increased likelihood of many topics to be repeated and similar comments about the same issues sent. The search procedures become difficult to implement in these cases because the content can be distributed along the forum's categories and many posts. In other cases, the user does not know the right word for searching.

In many cases, the sort of mechanisms available are not valid to find the interesting information and manage the forum. Normally, criteria to be used for sorting are date, user, category, and message title. But the results can be numerous and do not contain the required information.

Based on the foregoing observations, we assert that one of the most important research challenges for huge online courses and communities consists in proposing effective methods for enabling management of the forums by teachers and monitoring its content by students. This is the main problem addressed in the research presented in this work. The relevance of this challenge for ODL platforms is evident since more and more users going online are demanding this type of education. They will expect and require efficient and simplified access to information discussed in the forums.

The proposed method is based on applying NLP techniques for performing the forums analysis. The generated results will allow easier tracking and monitoring the online learning communities by users. The related work on this issue which will be examined more closely in the next section.

Related Work

The following subsections discuss the state-of-the-art aspects related to this research. A final subsection is added, which summarizes our contributions to previous work.

Related Work About MOOCs and ODL Management

As mentioned in the previous section, the management of MOOCs or ODL platforms is a very challenging issue. In these courses, forums are the key communication tool. There are studies which reveal that MOOCs rely mainly on discussion forums for interaction among students (Coetzee et al., 2014). This study carries out experiments with more than a thousand students on the "edX platform" (https://open.edx.org/) to conclude how forum design affects student activity and learning outcomes. They introduce a reputation system, which gives points to those students who make useful posts. Unfortunately, contrary to expectations, those reputation systems, apart from improving the forum experience, have no significant impact on grades, dropouts, etc. This fact, unequivocally, suggests that forums are essential for MOOCs, and other tools like reputation systems can improve the forum practice, but even more techniques are required in order to improve student outcomes and community formation.

Some methodology is required in order to improve the usability of the forums. In this way, the use of technology for enabling management is discussed in several works (Soledad Ramirez, Rivera, & Garcia, 2014). They conclude that further research lines are needed to enhance the role of technology in this challenge. Next, some approaches are cited in order to show how the technology is being used to enhance forum management.

There are proposals focused on developing a framework to classify MOOC discussion forum posts into a manageable number of categories (Stump, DeBoer, Whittinghill, & Breslow, 2013). Future innovations, such as Natural Language Processing (NLP), could also improve the discussion forum data.

Other researchers are using semantic filtering technologies for building MOOCs' management systems (Zhuhadar, Kruk, & Daday, 2015). Semantic technologies support more flexible information management than that offered by the current MOOCs' platforms and this is crucial for the success of any forum management.

There is other research focused on proposing systems for providing automatic mechanisms to manage the posts. For example, the re-grading system (Renz et al., 2014). This proposal presents the concept of automated re-grading with an equality approach in order to prove the meaning of transparent communication. This finding provides a guide for handling re-grading issues in vast learning environments. Other studies of this type propose systems for seeking events of the users as a means of management (Wong, Pursel, Divinsky, & Jansen, 2015). In this proposal, the system uses a keyword taxonomy approach in order to analyze a vast amount of MOOC forum data with the goal of identifying types of learning interactions and relations occurring in the forums.

Finally, a database system (such as MongoDB) is one of the most used technological tools for managing forums because of the flexibility to change information and perform further data analysis (Sarasa-Cabezuelo & Sierra-Rodríguez, 2014). Thanks to these systems, depth analysis can be made to obtain several pieces of information; for example, to identify the students who are likely to drop out of a MOOC (Tang, Xie, & Wong, 2015).

Related Work About Forum Analysis

Because of the importance of forums in ODL platforms, numerous studies and efforts were made to address issues about forum analysis, how it works, and implications for the development of courses.

There are two main research areas on forums analysis: (a) studies on forum structure, users' interactions, and types of students' and teachers' interventions; (b) studies focused on the content analysis of messages. Regarding the first set of studies (a), there are works aimed at understanding how students and teachers are using the forum and how they meet its expectations for learning and training (Shea & Bidjerano, 2009). The procedures consist of comparing instructor and student participation rates, reviewing the role of the instructors, and analyzing the users' interactivity to derive the learning outcomes and patterns of interactions (Onah et al., 2014; de Laat, Lally, Lipponen, & Simons, 2007; Suh & Lee, 2006). This kind of study provides qualitative and quantitative measures in order to find items that make the course more attractive (Suh & Lee, 2006; Swan, Shea, Fredericksen, Pickett, Pelz, & Maher, 2000), and help to improve their management. Therefore, from the level of participation and interaction of the discussion group, the researchers are trying to obtain the students' motivation (Baxter & Haycock, 2014; Yang, Sinha, Adamson, & Penstein Rose, 2013), performance (Romero, López, Luna, & Ventura , 2013), and also indicators to predict the degree of dropout of the online courses and communities (Yang et al. 2013).

The works focused on message content analysis (b) intend to find out information to assist users in meeting their learning or teaching objectives. These works mainly aim to know the influence of the forums from a

personal point of view in the student behaviour and their academic performance (Liu, Cheng, & Lin, 2013). The majority of forum posts deal with reporting questions, errors, and discussion about course material and organization. According to several analysis schemes (De Wever, Schellens, Valcke, & van Keer, 2006), there is a wide variety of work related to content analysis of forum posts. Most of them are focused on understanding the effectiveness of online forums as learning platforms for innovating the educational practices of teachers (Chávez, Montaño, & Barrera 2016), facilitating the teaching process (Brace-Govan, 2003), and providing evidence of students' learning as an aid to student assessment (Premagowrie, Kalai Vaani, & Ho, 2014; Dennen, 2008; McKenzie & Murphy, 2000). This information helps teachers and forum administrators to know if critical thinking skills are developed and if learning really occurs. This information is very useful for managing the online courses and designing learning strategies due to many students resisting participation in the forums (Guzdial & Turns, 2000).

Concerning the methodology used in the previous research, the reviewed work has used a variety of techniques. Firstly, there is work in manually analyzing forum content by labelling posts by categories of interest (Stump et al., 2013; McKenzie & Murphy, 2000; Coursaris & Liu ,2009). Other work uses statistical research procedures for the post collected by means questionnaires or surveys to the users (Premagowrie et al., 2014; Shea & Bidjerano, 2009). In addition, it should be highlighted that the work that designs computational methods for automatic analysis of users and messages sent to the forum. These computational methods perform interactions analysis on communication structures using: data mining techniques (Thomas, 2012; Fan, Zhang, Dang, & Chen, 2013; Anbalagan, Kumar, & Bijlani, 2015), social network analysis (Erlin & Rahman, 2009; Yang et al., 2013; de Laat et al., 2007; Suh & Lee, 2006), and/or other artificial intelligence methods (Li & Wu, 2010; D'Mello, Olney, & Person, 2010; Cade, Copeland, Person, & D'Mello, 2008), which are detailed next.

Data mining aims to explore the posts extracted from forums in order to discover structures and to understand the dynamics of the community. There are several techniques for providing means to automatically index, search, cluster, and structure the posts by discovering a set of topics within the forum. In this way, the forums could be represented by the topics within them. These techniques are usually based on statistical topic models (Thomas, 2012) and classification and clustering algorithms (Fan et al., 2013; Romero et al., 2013). Social network analysis offers a method for mapping group interactions, communication, and dynamics. The methods are usually implemented using computer-assisted qualitative data analysis software such as NVivo (Hoover & Koerber, 2011). This method codes the contributions into units of meaning which are assigned to parts of messages based on semantic features. Finally, other artificial intelligence methods for modelling dialogues and forum structures are based on machine learning approaches such as K-means clustering (Wang, 2013), Support Vector Machine (Li & Wu, 2010), and Hidden Markov Models (D'Mello et al., 2010; Cade et al., 2008).

Related Work About the Automatic Analysis of Text

Several research lines are currently developed for automatic analysis of text, the aim of which is close to our proposal in this paper: Topic Detection and Tracking, Recommender or Recommendation Systems, and News Trackers. These lines are detailed next.

Topic Detection and Tracking (TDT) is a Defence Advanced Research Projects Agency (DARPA) sponsored initiative to investigate the state of the art in finding and following new events in a stream of broadcast news stories (Allan, Carbonell, Doddington, Yamron, & Yang. 1998). Mainly, the TDT study pretends to explore techniques for detecting the appearance of new topics in a stream of stories, and for tracking the reappearance and the evolution of them. TDT techniques are applied to Social Networks (e.g., in Rho Rahayu, & Trang, 2015) in real data sets. In Allan et al. (2005), the TDT cluster detection technology was deployed in a real world setting, and they detected several drawbacks to be solved regarding the incremental clustering along time, which is a key issue in forum analysis.

Regarding the Recommender or Recommendation Systems (RS), they help to determine which information has to be offered to individual consumers and allow users to quickly find the personalized information that fits their needs (Hu, Dai, Song, Huang, & Chen. 2015). Nowadays, RSs are ubiquitous in various domains and e-commerce platforms, such as recommendation of books at Amazon, music at Last.fm, movies at Netflix and references at CiteULike. In the same line, Collaborative filtering (CF) approaches are extensively investigated in the research community and widely used in industry. They are based on the naive intuition that if users rated items similarly in the past, then they are likely to rate other items similarly in the future (Goldberg, Nichols, Oki, & Terry, 1992). In the educational domain, the proposal by Winoto, Tang, & McCalla (2012) makes personalized paper recommendations for users in this domain, for example, when the RS helps a tutor or learner to pick relevant courses, programs, or learning materials (books, articles, exams, etc.), and the contexts include the user's learning goals, background knowledge, motivation, and so on. Similarly, the proposal in Sathick and Venkat (2015) implements an online recommender application in the career guidance of online learners who pursue their graduation in an open and distance learning environment. Their aim is to help users to attain semantic knowledge from heterogeneous web sources and to make decisions.

Finally, the News Trackers systems (NT) show the applications of previous techniques in the real world. For example, the *MemeTracker* (http://www.memetracker.org/) by Leskovec, Backstrom, & Kleinberg (2009) builds maps of the daily news cycle by analyzing around 900,000 news stories and blog posts per day from 1 million online sources, ranging from mass media to personal blogs. They track the quotes and phrases that appear most frequently over time across this entire spectrum. The collection of distinctive phrases that will act as tracers for memes are the set of quoted phrases and sentences found in the articles. This makes it possible to see how different stories compete for news and blog coverage each day, and how certain stories persist while others fade quickly. Currently, many News Tracker applications are being developed, mainly in online newspapers such as *NewsTracker* (http://www.yournewstracker.com/).

Findings and Contributions of the Proposal

After reviewing the previous work, some findings can be drawn that justify and summarize our contributions to the state-of-the-art:

- Forums are one of the most important success factors of ODL courses. The great majority of online courses and communities have an open discussion forum to exchange messages asynchronously among users. However, forums are not a good tool for education when many messages are

produced in a disordered and unstructured way. Therefore, the existence of a forum in the course does not mean that it will be used to learn effectively.

- The study of the forums can have an impact on how instructors transfer knowledge and the results can be used for improving student's participation, retention, and learning. However, no effective solutions have been proposed to facilitate automatic monitoring of massive forums and enable them to remain as good interaction and communication tools.

- Proposals based on the automatic analysis methods offer promising results in processing massive quantities of data posts. Some work has been done in social networks analysis and other online platforms; however, these techniques have not been extensively used in the context of educational courses with massive forums. Consequently, these methods are potentially useful for improving learning-teaching processes, but they are still relatively immature for education applications.

Our paper is focused on processing the threads and post of the forums of online learning communities to facilitate tracking and monitoring by users. As an academic contribution, we propose a method that groups and summarizes the threads and post of the forums according to the subject matter being studied in order to describe its content and evolution along the course. In this regard, we seek to enhance the benefits of the teaching-learning process for both the learners and the teacher, especially in MOOC courses and learning communities with forums that can produce hundreds of posts per week.

The proposed method performs an automatic analysis of the educational forums by means artificial intelligence techniques based on Natural Language Processing. As a scientific contribution, it improves *Topic Detection and Tracking* and *News Trackers* systems by selecting a set of phrases instead of a single topic per forum thread and by simplifying this set of phrases to perform the incremental clustering (instead of selecting just one cluster centroid) and to improve the computational efficiency and precision.

Automatic Processing of Online Forums

In this section, our proposal is fully explained in two subsections. The first subsection overviews our proposal, whereas the second subsection details it.

Overview

The proposed method is depicted in Figure 1. Firstly, the text corresponding to the program of the subject is processed by NLP tools (the methods *POS-tagging_and_partial_parsing* and *Semantic_enrichment* in

Figure *1*). Secondly, it is clustered (*Clustering* method) and the most relevant topics of each cluster are extracted (*Representative_topic_extraction*). Next, each new student post is processed in a similar way in order to obtain the clusters of the set formed by the original post and the replies to it. In this way, in each separated thread of posts, we accomplish the anaphora resolution of definite descriptions such as "the practice" to the reference of the original topic "practice 2 of the diamond construction," which are both included in the same cluster. The aim of the extraction of the most relevant topics (*ltopics_post*) is to

improve the computational efficiency of our proposal with regard the incremental clustering, as well as to discard these anaphoric references. In this way, we cluster again, only these relevant topics of the new post, with regard to all the previous relevant student posts (*ltopics_post* + *ltopics_all_post*). Finally, the list of the most relevant topics from all the posts is processed jointly (*ltopics_all_post* + *ltopics_subject*) with the subject topics in order to obtain the final list of relevant topics (*ltopics*) linked to the subject program topics. Next, this algorithm is explained in full detail.

```
ALGORITHM  Topic_detection_of_subject

INPUT      subject_program: text;     // All the textual information for the subject
           lposts: list_of_post;      // List of posts and answers to the post
           Ont: Ontology;             // Ontology of the subject
           WN: Semantic_resources;    // For example WordNet

OUTPUT     ltopics: list_of_lists_of_SS;     // Most relevant detected topics

VAR        lSS_subject, lSS_post: list_of_SS;
           ltopics_subject, ltopics_all_post, ltopics_post: list_of_lists_of_SS;
           post: text;  // Set of texts that relate the post and the replies to it
BEGIN
    lSS_subject = POS-tagging_and_partial_parsing(subject_program);
    lSS_subject = Semantic_enrichment(lSS_subject, Ont, WN);
    ltopics_subject = Clustering(lSS_subject);
    ltopics_subject = Representative_topic_extraction(ltopics_subject);

    For each post in lposts
           lSS_post = POS-tagging_and_partial_parsing(post);
           lSS_post = Semantic_enrichment(lSS_post);
           ltopics_post = Clustering(lSS_post);    // lSS_post is converted into a
                        // list_of_lists_of_SS: [a, b, c] → [ [a], [b], [c] ]
           ltopics_post = Representative_topic_extraction(ltopics_post);

           ltopics_all_post = Clustering(ltopics_post + ltopics_all_post);
           ltopics_all_post = Representative_topic_extraction(ltopics_all_post);
    End For
    ltopics = Clustering(ltopics_all_post + ltopics_subject);
    ltopics = Representative_topic_extraction(ltopics);
END ALGORITHM
```

Figure 1. Topic detection method for a subject's posts.

Technical Details of the Automatic Processing Proposal

The text processing is performed by POS-*tagging_and_partial_parsing* in this algorithm (Figure 1). Firstly, the text is POS-tagged in order to obtain terms with their lexical category, lemma, and morphological information: noun (n), determiner (det), adjective (adj), verb (v), preposition (prep), singular (sing), etc. After that, it is partial parsed to extract noun phrases (NP), prepositional phrases (PP), and verbal phrases (VP), whereas the chunks not included in these phrases are skipped in the parsing. These phrases are fully parsed because NPs can have nested structures such as PPs, appositions or relative clauses as in "error in the practice of the diamond construction." Moreover, coordinated NPs and PPs are parsed as in "practice 2 and 3." These phrases represent the "main concepts" involved in the text.

These main concepts are stored in a list (delimited between square brackets) in the sequential order that they appear in the text. This list contains *slot structures* (SS) extracted from the parser (Ferrández, Palomar, & Moreno, 1999), which store the morphological knowledge (in the structure "conc," such as number and gender), an identifier (marked as upper cases such as X), syntactic knowledge (e.g., the slot structures of nested phrases), the term as it appears in the text (e.g., "got") and the lemma (e.g., "get"). In the SS examples, the slot structures and lemmas are simplified with the aim of clarity. This list of SS is simplified in order to keep only the phrases to cluster (e.g., the skipped chunk structures or the noun phrases with pronouns are removed) and is also extended (*Semantic_enrichment*) with semantic information obtained of the subject ontology (*Ont*) and additional semantic resources (*WN* such as WordNet). In this way, semantic comparisons can be done such as synonymy or hyponymy.

The Clustering process used is based on a hierarchical clustering dendogram that receives a list of lists of SS (llSS), in which each individual list represents a cluster. The output will be a new list of lists with the resulting clusters. For example, the Figure 2 shows an example of an input llSS and the output resulting clusters *llSS_F* :

```
[ [ np(Good afternoon) ],                                                    SS1
  [ vp(got) ],                                                               SS2
  [ np(the following error, pp(in, np(checking, pp(of, np(practice 2))))) ],  SS3
  [ np(practice 2) ],                                                        SS4
  [ pp(in, np(diamond practice)) ],                                          SS5
  [ np(error, pp(in, np(the practice, pp(of, np(the diamond construction))))) ]  SS6
]
```

(a)

```
[ [ np(Good afternoon) ],
  [ vp(got) ],
  [ np(the following error, pp(in, np(checking, pp(of, np(practice 2))))),
    np(error, pp(in, np(the practice, pp(of, np(the diamond construction))))),
    np(practice 2),
    pp(in, np(diamond practice))
  ]
]
```

(b)

Figure 2. Clustering process example (a) Input *llSS;* (b) Output *llSS_F*.

The clustering algorithm stores a similarity value between each pair of SS in a square matrix of dimension the length of *llSS*. In this section, the SS# are taken from the example depicted in Figure 2.(a). The Figure 3 shows the matrix of similarity values used in the clustering process (the similarity value between SS1 and SS2 is the same as between SS2 and SS1, that is why only half of the matrix is used). The algorithm iterates in order to join those rows with the highest similarity until a threshold is reached (e.g., the *llSS_F* when the similarity values between clusters are lower than the threshold). For example, in the first iteration in Figure 3, SS4 (*practice 2*) and SS5 (*in diamond practice*) are joined into the same cluster. As SS4 and SS5 appear in the same thread, it is quite likely that the anaphora reference is resolved rightly (i.e., they refer to the same entity).

	SS1	SS2	SS3	SS4	SS5	SS6
SS1	1					
SS2	0	1				
SS3	0	0	1			
SS4	0	0	0.4	1		
SS5	0	0	0.3	0.8	1	
SS6	0	0	0.5	0.6	0.4	1

Figure 3. Matrix of similarity values used in the clustering process.

It is usual to apply a filtering process to the set of SS in order to filter some frequent and irrelevant expressions (e.g., "Good afternoon", "Hi", determiners, prepositions, etc.). In this way, the main concepts are simplified, leaving only the core information discussed.

The similarity information is stored in a vector of Real values, where each value corresponds to each term in both phrases and ranges between 0 (null similarity) and 1 (maximum similarity). The Real values are calculated according to a lexico-semantic measure, in which different possibilities are considered: identical lemma, stem of lemma (e.g., *practice* vs. *practices* is measured as 0.9), synonym (e.g., *practice* vs. *exercise* is measured as 0.8), hyponym (e.g., *practice* vs. *activity* is measured as 0.7), and lexical distance (e.g., using the Levenshtein (1966) distance, *practice* vs. *practise* is measured as 0.6 divided by the number of single-character edits, i.e., insertions, deletions, or substitutions, required to change one term by the other), each one with its corresponding similarity value. In addition, these values are updated when they are compared with each nested phrase.

Finally, the head phrase similarity is calculated as the summation of the vector values, each one multiplied by the *lexical category weight* that measures the importance of the appearance of the term in the phrase (e.g., a determiner must have a lower weight than a noun). In this paper, we have used the following lexical weights: determiner (0.1), preposition (0.2), adjective (0.6), common noun (0.8) and proper noun (1). This summation is divided by the summation of the lexical weights for each term, in order to obtain a similarity value between 0 and 1. Following the same process, the similarity value is calculated for each different nested phrase. In the cases where null similarity is obtained in a nested phrase, as in *pp(in unit 1)*, these nested phrases will be compared with the remaining phrases in the text in order to be clustered in a different group. Therefore, possible mistakes in partial parsing regarding to incorrect attachment of nested phrases are solved. For example, the phrase "Could you clarify the following doubt in the title of the 4th practice?" the parser returns the nested phrase: "np(the following doubt, pp(in, np(the title, pp(of, np(the 4th practice)))))"; whereas the first prepositional phrase could be clustered separately to the first noun phrase.

With the aim of clarifying the application of the proposed algorithm, several examples are reported in Table 1. These examples have been simplified by not considering the lexical weights and semantic relations in order to ease the similarity calculation. We should clarify that similarity(the new error; the new practice 2) = 0 because of the mismatching of the heads. Similarly, similarity(error, pp(in practice 2); error, pp(in the practice, pp(of the diamond)) = 1/1 + 2/4 + 0/3 = 1.5, where the value 2/4 corresponds to the similarity(in practice 2; in the practice), and 0/3 is assigned to the similarity(of the diamond; remaining phrases). However, there is one addend for each nested phrase in the summation of similarity(practice 2, pp(with errors)); error, pp(in the practice, pp(of the diamond))), because there is no similarity value between each nested phrase.

Table 1

Examples of Similarity Calculation

SS1	SS2	Similarity
np(practice)	np(practice 2)	1/2 = 0.5
np(practice 3)	np(practice 2)	1/3 = 0.3
np(the last practice 2)	np(practice 2)	2/4 = 0.5
np(the last practice 2)	np(the difficult practice 2)	3/5 = 0.6
np(the new error)	np(the new practice 2)	0
Complex phrases with nested phrases		
np(practice)	np(the practice, pp(of the diamond))	1/2 + 0/3 = 0.5
np(diamond)	np(the practice, pp(of the diamond))	1/3 + 1/3 = 0.6
np(error, pp(in practice 2)	np(error, pp(in the practice, pp(of the diamond))	1/1 + 2/4 + 0/3 = 1.5
np(practice 2, pp(with errors))	np(error, pp(in the practice, pp(of the diamond)))	2/3 + 1/2 + 1/3 + 0/3 = 1.5
np(practice, pp(of diamond construction))	np(diamond construction practice)	3/3 + 2/3 = 1.6
np(the new error, pp(in the diamond construction))	pp(in the new practice, pp(of the diamond))	0 + 2/4 + 0 = 0.5

The *Representative_topic_extraction* process in

Figure *1* is based on Information Retrieval (IR) techniques, which from a user information need regarding a collection of information resources, its aim is to obtain these information resources sorted by relevance to the information need. IR is based on assigning weights to each term in the collection that measure the indication of relevance to future user queries. The relevance of each cluster is obtained from the summation of the weights of its terms, and the most relevant phrases in the cluster are obtained in the same way.

Case Study

In this section, we present an illustrative case study. We have carried out a set of experiments to validate our proposal. The subject on which we have analyzed the forums is *Fundamentals of Programming* (code: 71901020), taught in the first semester of the first course in the Bachelor in Computer Science degree. This degree is taught at the Spanish Open University (National University for Distance Education in Spain). The forums on this subject are divided into eight categories (as shown in Table 2). Initially, the system administrator, according to the recommendations of the teachers, defines the forum categories. When a student creates a new thread, he or she has to decide in which forum category the thread will be added. In the current academic year (2015/16) there are 586 students enrolled in the subject. Throughout this course, 926 posts have been written by the students with an average 61.7 posts per week. This fact shows the volume of information that must be processed by the teachers and students. If the advisor teaches several courses (of a similar number of students) then he or she will spend a lot of time in academic management of forums. Students have the same problem of finding the most relevant topics for each subject. In a standard academic course, the students have enrolled in eight subjects on average. That is, at the end of the course, the size of

all the forums can be of the order of many thousands of posts. So that, they must to review a lot of messages in order to find the right answers.

In the case of MOOCs with a larger number of students enrolled, the forum can become an unmanageable tool for finding valuable information and handle the students' posts. In this case, the forums may have ceased to be useful.

Table 2

Forum Categories in Fundamentals of Programming

Categories	Threads	Posts
Teaching staff	14	40
General questions	25	95
Students	31	270
Practices	73	319
Programming environment	20	108
C +/- Language	18	80
Tutorial coordination	3	11
Tutorial group	1	3
TOTAL	**185**	**926**

As previously mentioned, we propose an automatic information extraction method. The main objective of the method is to detect the different topics that students speak in the same forum category and classify them later. Consequently, common issues will be detected in the same category and, in the future, we will obtain the "self-managed" forums. The extraction method applied to one forum category can be extended to all forum categories in order to obtain all the topics on the subject.

In our example, we will focus on the "Practices" forum category (which contains 319 posts). The processing of the threads and posts will consist on the previously described three main phases: (1) lexical-morphological analysis, parsing and semantic enrichment of the post; (2) clustering and post relevant topic extraction; (3) clustering and thread relevant topic extraction.

The input of the processes acting on the posts is the set of threads to study. In this way, we can analyze different sets of threads; for example, we can study the threads of the same forum category or the threads between different dates (i.e., from the beginning up to mid semester) or it is also possible to analyze all the threads with the same title (i.e., practice delivery). Each post is processed individually. The output of the first phase is the post which contains lexical, morphological, syntactic, and semantic information. It is important to emphasize that linguistic phenomena resolution (such as definite description or anaphora resolution) has been applied.

The input of the second phase (the clustering procedure of a post) is a list of lists containing the partial parsing enriched with semantic information (each list is a phrase with partial parsing of the post). The

similarity matrix (between each pair of phrases) described in section 3 is calculated. After that, a list of clusters is obtained: (1) [*np*(Practice delivery), *np*(the practices 1,2 and 3), *vp*(delivered), *np*(the 4th practice)]; (2) [*np*(Good afternoon)]; (3) [*vp*(am watching)]; etc. Subsequently, clusters are ranked according to their relevance using IR techniques. Furthermore, phrases inside a cluster are reordered according to their relevance (the length of the phrase can be used as measure of relevance). Figure 4 shows the new cluster ranking: (1) [*np*(the practices 1,2 and 3), *np*(the 4th practice), *np*(Practice delivery), *vp*(delivered)]; (2) [*pp*(to *np*(the advisor)), *pp*(by *np*(the advisor))]; etc.

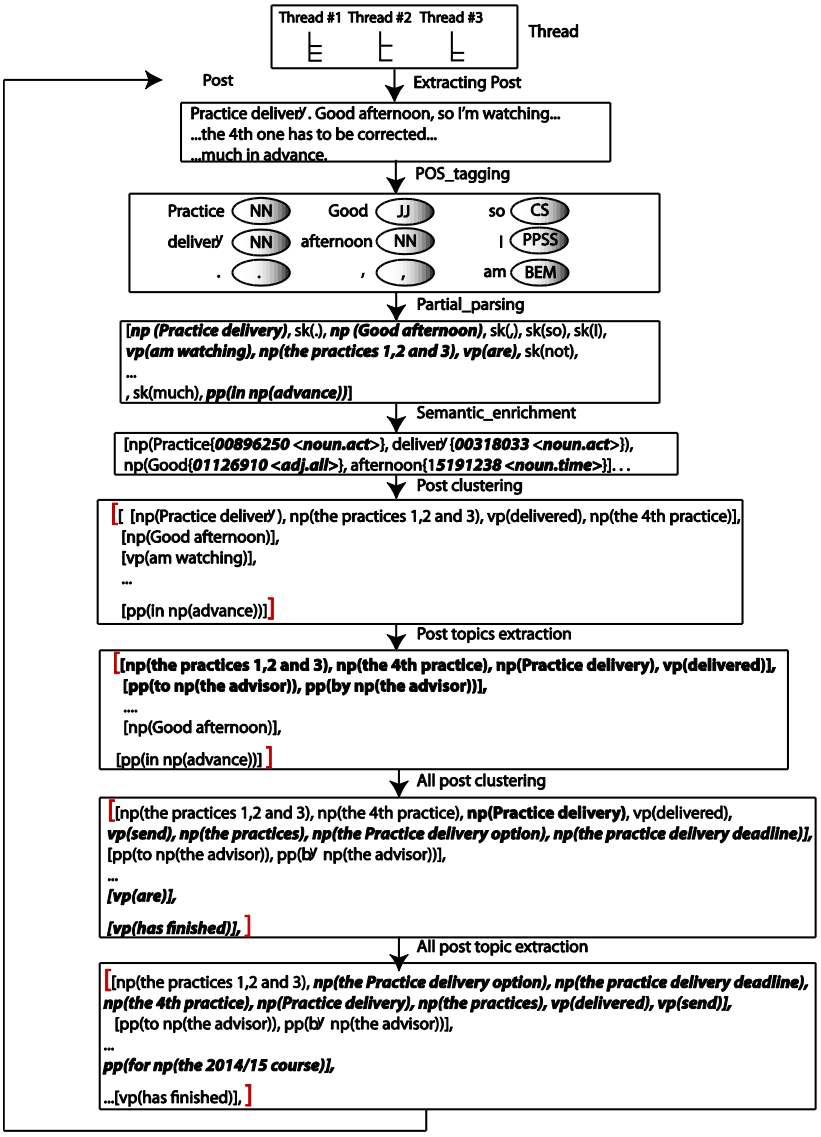

Figure 4. General process of the post.

To conclude the analysis of a thread we have to combine the clusters obtained from the new post with clusters of previous posts (phase 3). The process is similar to the one described above, by calculating the similarity matrix among each pair of phrases. With this process: (1) clusters could be enriched with phrases

(according to their similarity) from previous posts; (2) the number of clusters could be increased with the inclusion of the previous ones. The latter process (all post topic extraction) ranks again all clusters according to their relevance. Moreover, as we previously mentioned, a filtering process should be applied to the set of SS/clusters in order to filter some frequent and irrelevant expressions (e.g., "Good afternoon", "are," or "the advisor").

As it discussed along the paper, one of the main benefits of this work is the automation of the huge amount of posts included in the forums. The results of this study show the importance of our proposal because it is possible to display the main topics in a clean, orderly, and classified way with the aim to enrich the student-teacher interaction taking into account the contribution of knowledge among all the information included in the forums. In Figure 5 the main topics of the "Practices" forum category that have been detected by the system are shown.

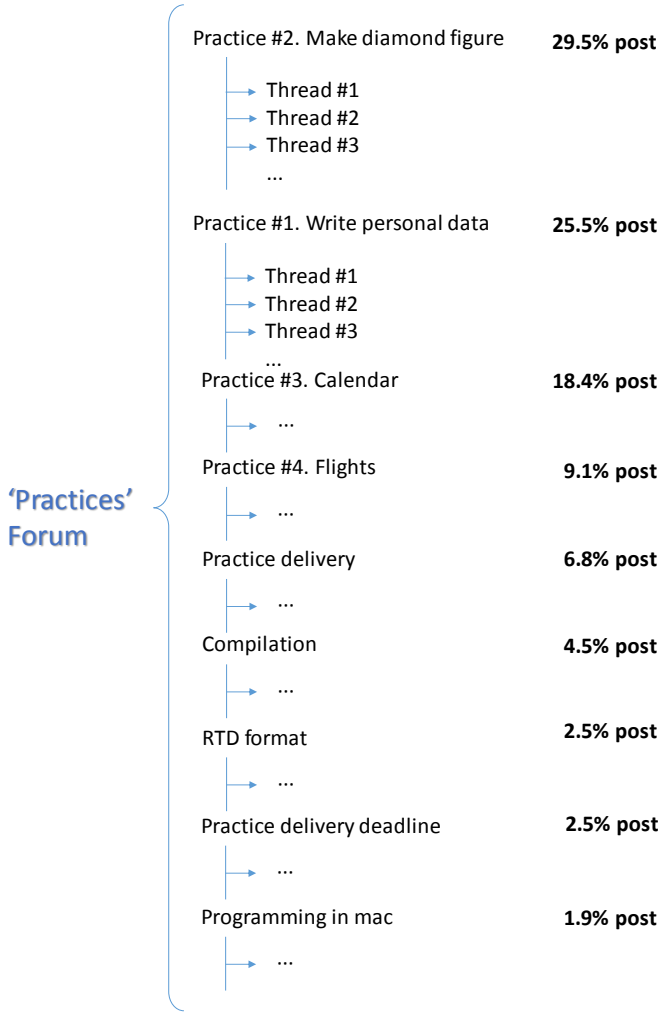

Figure 5. The main topics of the "Practices" forum category extracted by the system.

We can distinguish nine main topics ranked by the number of threads related to each topic. For instance, the most important topic is "Practice 2. Make diamond figure" with a percentage of 29.5% of the total

threads in this forum category; by contrast, “Programming in mac” is the least relevant topic with 1.9% of the threads.

The topics extracted that are shown in Figure 5 can change each year according to the development of the course. These results allow the teaching staff to understand what is discussed in the online course and focus their work on providing explanations about the most important issues. In addition, the student could identify much more quickly and efficiently with the topic they are looking for. For example, only 6 posts of 319 deal with “programming in mac.” The student with a question about mac issues would have to read most of the post or create a new thread with the issue, although it has already been discussed. In contrast, after the forum is processed, the student can identify and read the topics about mac easily.

Conclusions and Future Research

The ODL platforms in general, and MOOCs especially, have become very popular over the past few years. Forums are a central communication tool in many online educational platforms. These courses rely mainly on discussion forums for interaction among students. However, the advantages that this tool provides are very often attenuated due to several problems which are very demanding and time consuming. In these cases, forums do not support learning as expected because many messages may be produced, especially when they are posted in a disordered and unstructured way.

Numerous studies have been performed to look for information about students and other aspects of forum operation to improve management and learning effects. Nevertheless, no effective solutions have been proposed to facilitate automatic management and monitoring of massive forums. Automatic processing methods are potentially useful for this issue, but they are still relatively immature for education applications.

In this paper, we have presented a computational method to facilitate the tracking and monitoring of forums generated by online learning courses and communities. In this way, it enables the teaching-learning process, especially in huge online courses such as MOOCs, in which vast volume of information is generated by their forums. The proposed method allows a reduction of time spent by teachers with academic management and a better use of forums by students.

This approach performs an automatic analysis of the educational forums by means of Artificial Intelligence techniques based on Natural Language Processing: lexical, syntactic, and semantic analysis; clustering; and information retrieval. As an academic contribution, our paper proposes a tool that extracts the main topics discussed in the forums according to the subject matter being studied in order to describe its content and evolution along the course. Based on this information, many actions can be done: for example, the system can automatically restructure the forum categories according the main themes addressed, the teacher staff can identify what are the hot issues to provide useful explanations about them, and the students can identify easily the thread where to look for their answer or to post new comments. In this way, this information performs tracking the forums and the online course more effectively both for students and instructors. As a scientific contribution, it improves *Topic Detection and Tracking* and *News Trackers* systems by selecting

a set of phrases instead of a single topic per forum thread and by simplifying this set of phrases to perform the incremental clustering. It allows that clustering decisions are not irrevocable, and consequently improves the computational efficiency and precision. Moreover, we do not need to keep a vector of preferences for each user (teacher or learner) as it occurs in *Recommendation Systems*.

The implications of the results for ODL courses are evident. This research proposes innovative technologies for enabling smart education and provides a valuable contribution to open and distributed learning. The importance of an efficient management of the forums can help students to better understand the subject and finish their courses. This is one of the main challenges of MOOC platforms. Note that a difficult and tedious forum could easily invite students to drop out the course.

This proposal has been put in place through a real case study that shows promising results, although as future work we plan to perform a thorough evaluation in MOOCs, which will show its expected benefits. Moreover, we plan to devise different profiles of the method, distinguishing between the teacher and student profile, and its integration with the academic progress of the course (e.g., showing indicators of performance of the student in the course).

Acknowledgements

This paper has been partially supported by the MESOLAP (TIN2010-14860), GEODAS-BI (TIN2012-37493-C03-03), DIIM2.0 (PROMETEOII/2014/001) and RESCATA (TIN2015-65100-R) projects from the Spanish Ministry of Education and Competitiveness.

References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., & Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Landsdowne, VA.

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., & Amstutz, P. (2005). Taking topic detection from evaluation to practice. Annual Hawaii International Conference on System Sciences - Track 4 – 04, 1-10.

Anbalagan, R., Kumar, A., & Bijlani K. (2015), Footprint model for discussion forums in MOOC. Second International Symposium on Computer Vision and the Internet, *Procedia Computer Science, 58*, 530–537.

Baxter, J. A., & Haycock, J. (2014), Roles and student identities in online large course forums: Implications for practice. *International Review of Research in Open and Distance Learning, 15* (1), 20-40.

Brace-Govan, J. (2003). A method to track discussion forum activity: the Moderators' matrix. *Internet and Higher Education, 6*, 303-325.

Cade, W. L., Copeland, J. L., Person, N. K., & D'Mello, S. K. (2008). Dialogue modes in expert tutoring. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. Berlin: Springer-Verlag, 470-479.

Chávez J., Montaño, R., & Barrera, R., (2016). Structure and content of messages in an online environment. *Computers in Human Behavior, 54* (C), 560-568.

Coetzee, D., Fox, A., Hearst, M., & Hartmann, B. (2014). Should your MOOC forum use a reputation system? *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work &Social Computing*, 1176–1187.

Coursaris, C. K. & Liu, M. (2009), An analysis of social support exchanges in online HIV/AIDS self-help groups. *Computers in Human Behavior, 25*, 911–918.

De Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007), Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning, 2*(1), 87-103.

Dennen, V. P. (2008). Looking for evidence in learning: Assessment and analysis methods for online discourse. *Computers in Human Behavior, 24*, 205-219.

De Wever, B., Schellens, T., Valcke, M., & van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education, 46*, 6-28.

D'Mello, S., Olney, A., & Person, N., (2010). Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining, 1*, 1–37.

Erlin, N. Y. & Rahman, A. A. (2009). Students' interactions in online asynchronous discussion forum: A social network analysis. *International Conference on Education Technology and Computer*, 25-29.

Fan, L., Zhang, Y., Dang, Y., & Chen, H. (2013). Analyzing sentiments in Web 2.0 social media data in Chinese: experiments on business and marketing related Chinese Web forums. *Information Technology and Management, 14*(3), 231-242.

Ferrández, A., Palomar, M., & Moreno, L. (1999). An empirical approach to Spanish anaphora resolution. *Machine Translation, 14*(3/4), 191-216.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM, 35*(12), 61-70.

Guzdial, M. & Turns, J. (2000). Effective discussion through a computer-mediated anchored forum. *The Journal of Learning Science, 4*, 437-469.

Hew, K. F. & Cheung, W. S. (2014). Students' and instructors' use of Massive Open Online Courses (MOOCs): motivations and challenges. *Educational Research Review*, 12, 45-58.

Hoover R. S. & Koerber M. L. (2011). Using NVivo to answer the challenges of qualitative research in professional communication: Benefits and best practices. *IEEE Transactions on Professional Communication, 54*(1), 68-82.

Hu, G.-N., Dai, X.-Y., Song, Y., Huang, S.-J., & Chen, J.-J. (2015). A synthetic approach for recommendation: Combining ratings, social relations, and reviews. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, 24*, 1756-1762.

Lentell, H. & O'Rourke, J. (2004). Tutoring large numbers: An unmet challenge. *International Review of Research in Open and Distance Learning, 5*(1), 1-17.

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings for the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France*, 497-506.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals, Tech. Rep. 8.

Li, N. & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems, 48*, 354–368.

Liu, E. Z.-F., Cheng, S.-S., & Lin, C. H. (2013). The effects of using online q&a discussion forums with different characteristics as a learning resource. *Asia-Pacific Education Research, 22*(4), 667 – 675.

Lucas de Azevedo, V. & Borges, M. (2015). More collaboration, more collective intelligence. *International Journal of Knowledge Society Research (IJKSR), 6*(3), 1-18.

Lytras M. D., Mathkour, H. I., Abdalla, H., Wadee, A.-H., Yanez-Marquez, C., & Siquera, S. W. M. (2015). An emerging - Social and emerging computing enabled philosophical paradigm for collaborative learning systems: Toward high effective next generation learning systems for the knowledge society. *Computers in Human Behavior, 51*, 557-561.

Masud, M. (2015). Knowledge update in collaborative knowledge sharing systems. *International Journal of Knowledge Society Research (IJKSR), 6*(3), 19-31.

McKenzie, W. & Murphy, D. (2000). I hope this goes somewhere: Evaluation of an online discussion group. *Australasian Journal of Educational Technology, 16*(3).

Mora, H., Signes Pont, M. T. & DeMiguel Casado, G. (2014). Information search habits of first year college students. *International Journal of Knowledge Society Research (IJKSR), 5*(4), 26-34.

Mora H., Signes Pont, M. T., DeMiguel Casado, G., & Gilart Iglesias, V. (2015). Management of social networks in the educational process. *Computers in Human Behavior, 51*, 890-895.

Onah, D. F. O., Sinclair, J. E., Boyatt, R., & Foss, J. (2014), Massive open online courses: Learner participation. *Proceedings of the 9th International Conference of Education, Research and Innovation (iCERi)*, Seville, Spain.

Premagowrie, S., Kalai Vaani, R. & Ho, R. C. (2014) Online forum: A platform that affects students' learning? *American International Journal of Social Science, 3*(7), 107-116.

Ramesh, A., Goldwasser, D., Huang., B., Daumé, H., III., & Getoor, L. (2014), Understanding MOOC discussion forums using seeded LDA. *Proceedings of 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications*.

Renz, J., Staubitz, T., Willems, C., Klement, H., & Meinel, C. (2014, April). Handling re-grading of automatically graded assignments in MOOCs. *IEEE Global Engineering Education Conference, EDUCON*, 408–415.

Rho, S., Rahayu, W., Trang, U. (2015). Advanced issues on topic detection, tracking, and trend analysis for social multimedia. *Advances in Multimedia, 2015*, 1-2.

Romero, C., López, M. I., Luna, J. M., & Ventura, S., (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458-472.

Sarasa-Cabezuelo, A., & Sierra-Rodríguez, J.-L. (2014). Development of a MOOC management system. *Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality*, 155–162.

Sathick, J., Venkat, J. (2015). A generic framework for extraction of knowledge from social web sources (social networking websites) for an online recommendation system. *International Review of Research in Open and Distributed Learning, 16*(2), 247-271.

Shea, P. & Bidjerano, T. (2009). Community of inquiry as a theoretical framework to foster "epistemic engagement" and "cognitive presence" in online education. *Computers & Education, 52*(3), 543-553.

Sicilia, M.-Á., Lytras, M. D. (2005). The semantic learning organization. *The learning organization 12*(5), 402-410.

Soledad Ramirez, M., Rivera, N., & Garcia, A. (2014). MOOC learning: Challenges and opportunities of using team teaching. *Proceedings of the 7th International Conference of Education, Research and Innovation (iCERi2014)*, 5751–5756.

Stump, G. S., DeBoer., J., Whittinghill, J., & Breslow, L. (2013). *Development of a framework to classify MOOC discussion forum posts: Methodology and challenges*. Cambridge, MA: The Teaching and Learning Laboratory.

Suh, H. J. & Lee, S. W. (2006). Collaborative learning agent for promoting group interaction. *ETRI Journal, 28*(4), 461-474.

Swan, K., Shea, P., Fredericksen, E., Pickett, A., Pelz, W., & Maher, G. (2000). Building knowledge building communities: Consistency, contact and communication in the virtual classroom. *Journal of Educational Computing Research 23*(4), 389–413.

Tang, J. K. T., Xie, H., & Wong, T.-L. (2015). A big data framework for early identification of dropout students in MOOC. In J. Lam, K. K. Ng, S. K. S. Cheung, T. L. Wong, K. C. Li, & F. L. Wang (Eds.), *Technology in education. Technology-mediated proactive learning: Second international conference, ICTE 2015*, Hong Kong, China, July 2-4, 2015, Revised Selected Papers (pp. 127-132). Germany: Springer Berlin Heidelberg, 127-132.

Thomas, S. W. (2012). Mining software repositories with topic models (Technical Report No. 2012-586). School of Computing, Queen's University. Retrieved from http://cs.queensu.ca/~sthomas/data/Thomas_2012_TR2012-586.pdf

United Nations Educational, Scientific, and Cultural Organization (UNESCO). (2002). Open and distance learning: Trends, policy and strategy considerations. Retrieved January 15, 2016 from: http://unesdoc.unesco.org/images/0012/001284/128463e.pdf

Wang G., (2013). Research on hotspot discovery in internet public opinions based on improved *K*-means. *Computational Intelligence and Neuroscience*, 2013. doi:10.1155/2013/230946.

Winoto, P., Tang, T., & McCalla, G. (2012). Contexts in a paper recommendation system with collaborative filtering. *The International Review of Research in Open and Distributed Learning, 13*(5), 56-75.

Wong, J. S., Pursel, B., Divinsky, A., & Jansen, B. J. (2015). Analyzing MOOC discussion forum messages to identify cognitive learning exchanges. *Proceedings of the 2015 Annual Meeting of The Association for Information Science & Technology (ASSIST 2015)*, 1–10.

Yang, D., Sinha., T., Adamson, D., & Penstein Rose, C. (2013). Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. NIPS Data - Driven Education Workshop, 2013.

Zhuhadar, L., Kruk, S. R., & Daday, J. (2015). Semantically enriched Massive Open Online Courses (MOOCs) platform. *Computers in Human Behavior, 51*, 578–593.