An automated content analysis model for forum moodle using semantic similarity

A Tenriawaru
Department of Mathematics, Faculty of Mathematics and Natural Science, Halu Oleo university, Kendari 93231, Indonesia
atenriawaru36@gmail.com, andi.tenriawaru@uho.ac.id

Abstract. The use of e-learning in Indonesia’s educational environment is increasingly widespread. It requires an improvement in the method of evaluating learning outcomes in e-learning. Moodle is one of the most popular learning management systems in the world of education, but Moodle does not yet have the tools to evaluate learner activities in every Moodle activity. For example, Moodle doesn't have the tools to analyze the content of activity forums. This study proposes a model for analyzing message content in forums automatically. This model specifically aims to determine the relevance of the message to the course. The proposed model consists of three main processes: creating a dataset, detecting message relevance, and testing. Determination of the relevance of the message is based on the value of semantic similarity. Semantic similarity calculations are performed using Indonesian Wordnet and Levenshtein Distance Similarity. This model can help lecturers to detect which messages are relevant or worth and which messages are irrelevant or worthless toward the course. The proposed model is expected to improve the performance of evaluation systems of e-learning learning.

1. Introduction
The development of information technology is currently growing rapidly, this is encouraging those who are struggling in the world of education to develop learning methods. At this time, conventional learning methods have not been abandoned, but have been combined with e-learning methods, known as blended-learning.

Given the rampant implementation of e-learning in the educational environment especially universities in Indonesia, an evaluation tool for the quality of learning in e-learning needs to be developed. Various reasons for the importance of evaluating the quality of e-learning learning have been presented in previous studies [1-2]. There are several learning management systems (LMS) that can be used in developing e-learning learning, one of which is Moodle. Moodle provides many activities, such as quizzes, forums, assignments, chat, glossary, wikis, and lessons. However, Moodle does not yet provide tools to assess learner action on every activity. In the quiz activity, there are tools to assess answers to questions in the form of multiple choice and short answers, but Moodle does not yet have tools to assess the content of messages posted in forum activities.

There are several studies that have measured learner learning outcomes in e-learning by involving forum activities [3-6], but some of these studies only count hits or the number of learner actions without considering the contents of the message.
There are several studies that have measured learner learning outcomes in e-learning by involving forum activities [4]. Some of these studies only count hits or the number of learner actions without considering the contents of the message. Others have considered the contents of the message, but the forum message content assessment is still done manually, so measuring learner learning outcomes in e-learning is not yet fully automatic. To develop a model or automatic assessment tool on e-learning, tools are needed to assess the content of learner activities in the forum. Therefore, this research offers a model to detect the relevance of a message in forum Moodle to a course or discussion material.

2. Literature Review
As explained in the introduction, many studies have measured learner learning outcomes in e-learning learning. There are some studies which do not involve forum activities and some that involve forum activities.

Lara et al. used the number of hits for each Moodle activity to predict and classify learners in drop-out or non-drop-out clusters. Data used in the classification process include number of course accesses by learners in the week, number of different days of the week which the learner accesses the course, number of times that the learner has visualized the resource in the week. This research does not involve data from activity forums [5].

In 2012, Zafra and Ventura used assignment, forum and quiz activities to predict learner performance in a web-based education system. Forum activities have been involved, but only count the number of messages read, the number of messages posted, and the amount of time spent in forum activities. This research does not yet involve evaluating the contents of the message [6].

Shukor et al. used resource view actions, forum views, and forum posts to evaluate the quality of online learning through learner cognitive involvement. The level of cognitive involvement of learners is determined using content analysis and data mining methods. This research involves two types of data, namely learners' participation data and messages from the forum. Learners' participation includes the frequency of log-ins, resource views, discussion views, and add posts. The message data from the forum is a message content analysis, which is in the form of a message cognitive level. Although this research has focused on message content analysis, it is still done manually [7].

Romero et al predicted learner outcomes in the failure or success category, not predicting final scores in numerical form. In determining the success or failure category of learners, this paper uses classification and clustering. Clustering and classification are carried out based on three different types of analytic data from forum activities, namely Quantitative information, Qualitative information, and Social network information. Quantitative information is statistical information such as number of messages read, number of messages read posted, and time spent. Qualitative information is the score of the content of the messages. The score is given by the teacher based on the relevance of the message. Message rated from 0-3. Value 0 is given for messages that have no relevance to the course. A value of 1 is given for messages that provide limited information about the course, a value of 2 is given for messages to messages that provide adequate information about important topics. And a value of 3 is given for messages that provide very complete or precise information about difficult topics. The third data type, Social network information that uses questioning and responding relationships between learners. Although this research has involved content analysis of the message content of forum activities, the analysis is still done manually by the teacher [4].

3. The Proposed Conceptual Model
As described in the previous section, this paper is concerned with an initial investigation to develop an additional tool that will facilitate the e-learning to automatically analyse content of Forum message.

A framework is structured to detect the relevance of the message to the course. It consists of three stages, i.e. creating a dataset, detecting message relevance, and testing results. The design of the framework is shown in Figure 1.
3.1. Creating a dataset
At this stage, a dataset is created that will be used as training data and testing data. In this study, two datasets were created from messages in the Moodle forum. One dataset contains messages that are categorized as relevant and not relevant to the description of the forum by the annotators and one dataset that contains messages that are categorized as relevant and not relevant to its parents by the annotators. The annotators are provided with guidance documents to carry out annotations, which then manually annotate each message submitted, which are categorized as relevant and which are not. Figure 2 illustrates the process of creating a dataset.

The documents used in this study are messages obtained from forum activity data in courses that use Indonesian.

3.2. Detecting message relevance
This stage consists of three processes, i.e. the process of calculating semantic similarities between messages and forum descriptions, the calculation of semantic similarities between messages and its parent, the process of determining threshold values, and the process of determining the relevance of messages based on semantic similarity and threshold. The framework design is shown in Figure 3.
3.2.1. calculating semantic similarity

The semantic similarity calculation process is done by calculating the semantic similarity between the message and the forum description, and between the message and its parent using Indonesian WordNet and Levenshtein Distance Similarity. This process consists of six sub-processes, i.e. tokenization, Stop-word removal, stemming, POS tagging, making semantic similarity relative matrix, calculating the similarity value with the heuristic method. The following are the steps to calculate the Semantic \( \text{Sem}(K_1, K_2) \) equation between facts \( K_1 \) and \( K_2 \):

a. Stop-word removal: the process of deleting words that could potentially be a stop-word.

b. Stemming: the process of determining the basic word by removing the affixes that exist in a word that occur due to changes in the morphology of a word.

c. POS tagging: the process of labeling every word in a sentence.

d. Making Matrix \( R[m,n] \): the process of making Semantic Similarity Relative Matrix \( R[m,n] \) of for each pair of word sense with \( m \) is number of row and \( n \) is number of column. \( R[m,n] \) is the semantic similarity between the sense that best matches the words at position \( i \) of \( X \) with the most suitable sense of words at position \( j \) from \( Y \). So, \( R[m,n] \) is the weight of the relationship from point \( i \) to \( j \). If a word is not in the WordNet dictionary, then use Levenshtein Distance Similarity instead, by changing the similarity distance and outputting the lowest weight. The Levenshtein Distance Similarity calculation process can be indicated by the pseudo-code below:

```plaintext
for i from 1 to m:
    d[i, 0] := i
for j from 1 to n:
    d[0, j] := j
for j from 1 to n:
    for i from 1 to m:
        if s[i] = t[j]:
            d[i, j] := d[i-1, j-1]
        else:
            d[i, j] := min(d[i-1, j], d[i, j-1], d[i-1, j-1]) + 1
```

Figure 3 The framework of detecting message relevance
d[i, j] := \min(d[i-1, j] + 1, \\
               d[i, j-1] + 1, \\
               d[i-1, j-1] + 1)

e. Semantic similarity calculation: the process of calculating the semantic similarities \( \text{Sem}(K_1, K_2) \) between facts \( K_1 \) and \( K_2 \) from \( \text{Matrix } R[m,n] \). If \( \max(\text{row}_i) \) is the maximum value of row \( i \), \( \max(\text{col}_j) \) is the maximum value of column \( j \), \( n \) is the number of rows, and \( m \) is the number of columns, then the \( \text{Sem}(K_1, K_2) \) values calculated using the fast heuristic method with Equation (1).

\[
\text{Sem}(K_1, K_2) = \frac{\sum_{i=1}^{m} \max(\text{row}_i) + \sum_{j=1}^{n} \max(\text{col}_j)}{m \times n}
\]  

The process of calculating the semantic similarities \( \text{Sem}(K_1, K_2) \) can be indicated by the pseudo-code below:

```java
for (int i=0; i < m; i++)
{
    maxSim_i=0.0F;
    for (int j=0; j < n; j++)
        if (maxSim_i < simMatrix[i][j])
            maxSim_i=simMatrix[i][j];
    sumSim_i += maxSim_i;
}

for (int j=0; j < n; j++)
{
    maxSim_j=0.0F;
    for (int i=0; i < m; i++)
        if (maxSim_j < simMatrix[i][j])
            maxSim_j=simMatrix[i][j];
    sumSim_j += maxSim_j;
}

sim=(sumSim_i + sumSim_j)/(float) (m + n);
```

3.2.2. Determining threshold value

The next step is determining the threshold value. Threshold value is the value of the limit parameter determined to find out whether the message is relevant to the course or not. The threshold value is determined based on the value of semantic similarity between the message and the forum description, and between the message and its parent.

The scenario to determine the threshold value is by observing the semantic similarity values in the training data, then testing with several threshold values so that the right threshold value is obtained, which will improve the accuracy of the model.

3.2.3. Determining the relevance of messages

At this stage, checking and determining the relevance category of the message is carried out. Messages are categorized into relevant or irrelevant messages to forum description or its parent message based on the threshold value. If the semantic similarity value is greater or equal to the threshold then the message is concluded relevant to the forum or the parent. If the similarity value is smaller than the threshold, the conclusion is that the message is not relevant. This is shown in Equation 2. Based on the results of determining the relevance of the message to the forum and the parent, the relevance of the message to course will be determined.
6

3.3. Testing

Stages of testing will be carried out with scenarios to test the performance of the proposed framework. Testing is performed on testing data taken from previously created datasets.

4. Conclusion

An automated tool for determining the relevance of messages posted by learners to forums is essential for assessing learner learning through e-learning. With this tool, it can help teachers to detect which messages are relevant or have value toward the course and which message is only rubbish. This paper proposed a model for detecting the relevance of messages in forum Moodle toward the course. The proposed model consists of three main stages: creating a dataset, detecting message relevance, and testing. Making training data and testing data is done at the stage of creating a dataset. With training data, the criteria for relevant and irrelevant messages are determined. This determination is done at the detecting message relevance stage. Next, to ensure the accuracy and reliability of the model built, testing is done.

5. References

[1] Tenriawaru A, Djunaidy A and Siahaan D 2014 A new model of students participation measurement in e-learning systems based on meaningful learning characteristics: An initial investigation Technology, Informatics, Management, Engineering, and Environment (TIME-E), 2014 2nd International Conference on (IEEE) pp 96–9
[2] Tenriawaru A, Djunaidy A and Siahaan D 2016 Mapping Metric Between Meaningful Learning Characteristics and Moodle Activities Int. Rev. Comput. Softw. IRECOS 11 pp 1083–92
[3] Lara J A, Lizcano D, Martínez M A, Pazos J and Riera T 2014 A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA Comput. Educ. 72 pp 23–36
[4] Romero C, López M I, Luna J M and Ventura S 2013 Predicting students’ final performance from participation in on-line discussion forums Comput. Educ. 68 pp 458–72
[5] Lara J A, Lizcano D, Martínez M A, Pazos J and Riera T 2014 A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA Comput. Educ. 72 pp 23–36
[6] Zafra A and Ventura S 2012 Multi-instance genetic programming for predicting student performance in web based educational environments Appl. Soft Comput. 12 2693–706
[7] Shukor N A, Tasir Z, Zaid N and Ahmad N 2014 The Design of a Computer-Supported Collaborative Learning Environment that Promotes Interaction Teaching and Learning in Computing and Engineering (LaTiCE2014 International Conference on (IEEE) pp 47–51
[8] Simpson T and Dao T 2010 WordNet-based semantic similarity measurement - CodeProject