Neural Network Method in Text Message Categorization of Online Discussion

Erlin [1], Johan [2], Triyani Arita Fitri [3], Agustin [4] , Hamdani [5]

[1,2] Department of Informatic Engineering, Faculty of Computer Science, Institut Bisnis dan Teknologi Pelita Indonesia, Pekanbaru, Indonesia

[3,4,5] Department of Informatic Engineering, STMIK Amik Riau, Pekanbaru, Indonesia

Abstract

This paper presents research in neural network approach for text messages categorization of collaborative learning skill in an online discussion. Although a neural network is a popular method for text categorization in the research area of

machine learning, unfortunately, the use of neural network in educational settings is rare. Usually, text categorization by neural network is employed to categorize news articles, emails, product reviews, and web pages. In an online discussion, text categorization that is used to classify the message sent by the student into a certain category is often manual, requiring skilled human specialists. However, human categorization is not an effective way for a number of reasons; time- consuming, labor-intensive, lack of consistency in a category, and costly. Therefore, this paper proposes a neural network approach to code the messageautomatically. Results show that neural networks achieving useful classification on eight categories of collaborative learning skills in an online discussion as measured based on precision, recall, and balanced F-measure.

1. Introduction

Text categorization is a prominent method in the research into collaborative learning and active research area in information retrieval and machine learning. Text categorization refers to solve a problem to classify documents into a certain number of pre-defined categories based on their content.As the volume of a message in an online discussion increases and categorize the message into some classes is needed; thus, categorizing and coding a text message by a human is time-consuming, labor-intensive, and costly. Dumais [1] stated that human

systems' weaknesses are inconsistency in assigning category and needs to adapt to changing category structures.

The online discussion presents significant challenges to the existing text categorization technique. Online discussion messages are usually incomplete, error-prone, and poorly structured [2]. In an online discussion, text categorization that is used to classify the message sent by a student into a certain category is often manual, requiring skilled specialists. Therefore, how to use various computer technologies to auto-coding of a message is the subject of great research

Erlin
Email: erlin@lecturer.pelitaindonesia.ac.id

value. In recent years, research has shown that there has been extensive study and actively explored various machine learning for text document categorization and classification. Among these are Bayesian network classifier [3], *k*-nearest neighbor classifier [4] and decision tree [5]. These methods are conventional learning methods compared to those new approaches, however, it have simple algorithms and relatively high efficiency.

Recently, a number of researchers have been proposed some new approaches and models. For example, there are neural networks [6], support vector machines [7], fuzzy *k*-means [8] and maximum entropy models [9] that also have good results. Gabrilovich & Markovitch [10] have been proved that support vector machines (SVM) is one of the best algorithms for text categorization. Meanwhile, Yu, ben Xu & hua Li [11] have argued that neural network (NN) also a popular categorization method that can handle linear and non linear problems for text categorization, and both of linear [12] and non linear [13] neural network classifier can achieve good result. Unfortunately, the use of neural network in an educational setting is rare. Most of neural network is used to categorize news article, emails, product reviews, web pages and so on. Hence, our research focused on automated text message categorization in online discussion using neural network that can be trained on large corpus of messages and associates a high score to pairs of message-category that appear in the corpus data.

The rest of the paper is organized as follows: In the next section, we describe the collaborative learning skill in online discussion. Section 3, explains the method of this research. The experiment is described in section 4. The experiment results are given in section 5. Finally, the conclusions are given in section 6.

2. Collaborative Learning Skill

Many educators are really aware of the use of collaborative learning and information and communication technology (ICT) to facilitate the learning process for their students' benefit. Gokhale [14] has argued that collaborative learning is about groups of students, and groups of students and teachers, constructing knowledge together. In collaborative learning, students with various performance levels work together in small groups, and the students are responsible for one another's learning and their own. Hence, the success of one student helps other students to be successful.

Skill in learning collaboratively means knowing when and how to question, inform, and motivate one's teammates, knowing how to mediate and facilitate conversation, and knowing how to deal with conflicting opinions [15]. The collaborative learning skill category based on Soller's model, which is a modified version of McManus and Aiken's Collaborative Skills Network [16], is adopted for this research. In Soller's model, each conversation act is assigned a sentence opener indicating the act's intention. Students communicated through a sentence opener interface by initiating each contribution with one of the key phrases, conveying the appropriate dialogue intention.

The weakness of the sentence opener approach is limitations in using ideas or thinking that will be delivered in a discussion. Each student communicated by sentence opener first before posting message on the online discussion. Therefore, this research is designed for automated text categorization; hence students feel free to deliver their idea without being limited by the sentence opener that has been set previously by the system.

Table 1. Definitions and examples of collaborative learning conversation skills

No	Category	Description	Examples

1	Acknowledge	:	Inform peers that you read and/or appreciate their comments. Answer yes/no questions.	Thanks friends for the response that you all give for this question 1. Ok, I can accept this answer and thanks for your cooperation.
2	Discuss	:	Reason (positively or negatively) about comments or suggestions made by team members.	I agree because the example from our lecturer is very clear. However, I think we can more focusing on the content.
3	Inform	:	Direct or advance the conversation by providing information or advice.	According to my answer before, I have give the explanation about the errors or mistakes that have been occur based on Minicase 2. At last I know what functional information system is and also provide examples.
4	Maintenance	:	Support group cohesion and peer involvement.	Excuse me, but I don't think that's what you meant at all. Sorry if I was a bit technical answer.
5	Mediate	:	Recommended an instructor intervene to answer a question.	We should ask our lectures when we must submit our assignment.
6	Motivate	:	Providing positive feedback and reinforcement.	Very good, your idea is brilliant. It is help us to solve our problem. Do not give up, you have time to do your assignment until next week.
7	Request	:	Ask for help/advice in solving the problem, or in understanding a team-mates comment.	Could you help me to find an application for doing this task? Please answer the question below and elaborate your answer based on Chapter 7.

Table 1 offers brief descriptions of thesub-skill categories of Soller's model. We take eight sub-skill categories; request, inform, motivate, task, maintenance, acknowledge, discuss and mediate; to be implemented and tested using neural network.

3. Method

In text categorization method, documents are to be classified into a certain number of categories. In this research, each document represents a message in an online discussion. The message of an online discussion forum from one subject SCK3433-02 2008/2009: Management of Organisation Information Systems held on Moodle as learning management systems (LMS) in e-learning is examined for one discussion topic. This topic is chosen because it has highest replies than other topics. There are 29 students who completed the topic of discussion. The total numbers of messages in the online discussion (that are replies to somebody's message) are 394 messages.

A document from an online discussion is composed of the discussion subject, user name, user ID, date, time, and message contents. In our approach, we use only the contents of a message. The content of message is essential to analyze the dynamics of the forum and what kind of feedback from one another. It is also used to gather information about the quality of the online discussion. Moreover, the text categorization approach codes the content of messages according to the message type. A central idea in text categorization that the many words of the documents are classified into specific message categories.

The problem of text categorization may be formalized as the task to approximate an unknown classification function $\Phi : d \times c \rightarrow$ Boolean defined as:

$$\Phi\,(d,\,c) = \begin{cases} true, if\ d \in c \\ false, otherwise \end{cases}$$

Where d is a set of document, c is set of categories, for any pair (d, c) of document and category [17].

Auto-coding based on Neural Network (ACNN) model is shown in figure 1. First of all, students' interaction through online discussion as a means of sharing knowledge and solving a problem by posting their idea or solution in the text form. All text of this forum as known as corpus data will be categorized into 8 categories using neural network approach. As depicted in the figure, there are three stages in text categorization are applied: pre-processing, dimensionality reduction and classification.

3.1. Pre-Processing

Classifier or algorithm cannot be directly interpreted the text. Documents should first be transformed into a representation suitable for the classification algorithms to be applied. In order to transform a text into a feature vector, pre-processing is needed. This stage consists of identifying feature by feature extraction and feature weighting. The main goal of feature extraction is to transform a message from text format into a list of words as feature set, easier to be processed by neural network algorithms. This includes tokenization, stop word removal and stemming.

Tokenization is used to separate text into individual words. All upper case characters in the words are converted to lower case characters. Next, stop word removal to remove common words that are usually not useful for text categorization and ignored for later processing such as "are", "is", "the", "a", etc, based on a stop list for general English text. The remaining words then stemmed using the Porter's algorithm [18] to normalize words derived from the same root such as "computer", "computation", "computing" would end up into the common form "comput" after the applying the stemming process. After stemming, we merged the sets of word's stems from each of the 275 training documents and removed the duplicates. As a result are 1137 terms in the vocabulary is potential as feature set.

3.2. Dimentionality Reduction

The main difficulty in the application of a neural network to text categorization is the high dimensionality of the input feature space which is typical for textual data. This is because each term in the feature set represents one dimension in the feature space, as consequence, the size of the input of the neural network depends upon the number of stemmed-words after removed the duplicates. In order to reduce the dimension of the input, feature selection also called term space reduction (TSR) is needed to select, from the original set of term, that when used for document indexing, yield the highest effectiveness [19].

There are many kinds of document reduction techniques. Savio and Lee [13] introduce four kinds of methods designed to the dimensional of the feature space; document frequency (DF) method, Category Frequency-Document Frequency (CF-DF) method, product of term occurrence frequency (TF) and the inverse document frequency (IDF) as known as TF.IDF method and Principal Component Analysis (PCA). This research reduced the size of dimension by computing the document frequency (DF). Yang and Pedersen [20] have shown that with DF it is possible to reduce the dimensionality by a factor of 10 with no loss in effectiveness. This seems to indicate that the terms occurring most frequently in the corpus are the most valuable for text categorization.

3.3. Text Classifier

The neural network classifier must be trained before it can be used for text categorization. Training of the neural network classifier is done by the back-propagation learning rule based on supervised learning. In order to train the neural network, a set of training documents and a specification of the pre-defined categories the documents belong to are required.

In this research, three layer of back-propagation neural network is used which

consist of an input layer, a hidden layer and an output layer with the sigmoid function as the activation function. In the input layer, the number of input node is equal to the number of feature set after dimensionality reduction. In hidden layer, the number of hidden node affects the generalization performance. The

number of hidden node depends on the number of input node, output node and a constant ranging from (1, 10). In the output layer, the number of output node is equal to the number of pre-defined categories of text message categorization.
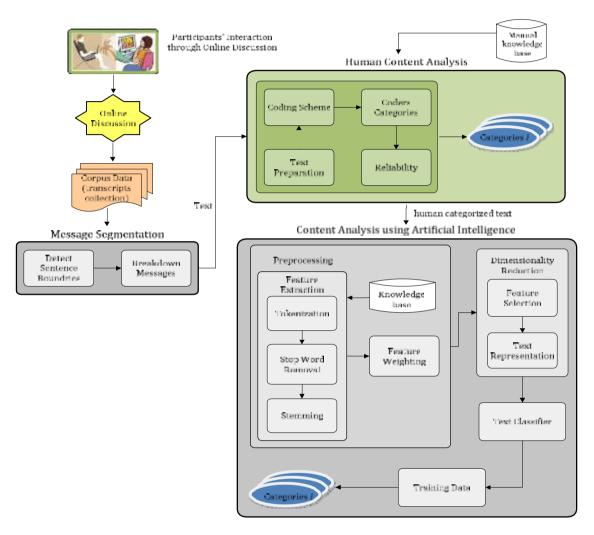

Figure 1. Framework for text categorization

4. Experiments

The messages are firs manually coded by human coder. The total number of messages that coded by human is 394. The list messages of online discussion categorized by two of human coders and calculated the reliability among them. The reliability of human coder is needed to determine how well the human coded the list of message based on coder training. The reliability test is conducted using multiple reliability coefficients such as percent agreement,

Scott's Pi (π), Cohen's Kappa (k) and Krippendorff's alpha (α). De Wever [21] has argued that reporting multiple reliability indices is of importance considering the fact that no unambiguous standards are available to judge reliability values. The coders do a sample exercise on other messages to familiarize themselves with the model. Two coders should do the analysis independently and have the results cross examined by one another. Thereliability value of human coders to

categorize the messages as follows: Percent Agreement=77.66%, π =73.49%; k=73.56% and α =73.52%.

Krippendorff [22] added that variable with Alpha as low as .667 could be acceptable for drawing tentative conclusions. The values of .667 also appropriate for Scott Pi and Cohen Kappa. The reliability values of categorization messages by human coders as shown in table 2.

Table 2. The reliability of message categories by human coder

N Agreements	N Disagreements	Percent Agreement	Scott Pi	Cohen's Kappa	Krippendorff Alpha
306	88	77.66%	73.49%	73.56%	73.52%

Armed with 394 human-code message, the neural network approach is ready to be trained. The total number of message is split into two parts. First part for training data and the second part for testing data our text categorization model. We used a set of 275 messages as training set for training the text classifier. For testing, a test set of 119 messages is used. In this research, we used 8 categories based on Soller's model. Each message of text is classified in the same category are presumed to have similar meaning. There are a total of 394 messages belonging to one of the 8 chosen categories. The distribution of messages among the 8 categories in the training set and the testing set is shown in table 3.

Table 3. No. of training set and testing set based on category

Name of Category	Number of Training Documents	Number of Testing Documents	Total
Acknowledge	40	15	55
Discuss	50	25	75
Inform	60	25	85
Maintenance	29	12	41
Mediate	2	1	3
Motivate	20	8	28
Request	44	20	64
Task	30	13	43
Total	**275**	**119**	**394**

In order to use a neural network learning approach, we first need to transform a text into a feature vector representation. Hence the feature extraction is needed. We created a program to combine the three phases of feature extraction in C++ language based on Porter's algorithm. The result can be seen in figure 2 and figure 3 that shows list words after stemming based on each message and list of potential words as a feature set from the whole message.

After stemming, we merged the sets of word's stems from each of the 275 training documents and removed the duplicates. As a

result are 11134 terms in the vocabulary is potential as feature set. After feature extraction phase that select the important terms, we should do term weighting of each word. There are various term weighting approaches studied in the literature. Boolean weighting is one of the most commonly used. In boolean weighting, the weight of a term is considered to be 1 if the term appears in the document and it is considered to be 0 if the term does not appear in the document.

The size of dimension is reduced by document frequency (*DF*) method. All potential features are ranked for each category based on the term occurs in the message. The top features for each category are chosen as its feature set. We choose 369 term as the neural network's input.

Figure 2. List of words after stemming

Figure 3. List of potential words as features set

The number of output nodes is equal to the number of pre-defined categories. According to the rule of hidden node selection:

$$N=\sqrt{n + m + a}$$

Where N is the number of hidden nodes, n is the number of input nodes, m is the number of output nodes and a is a constant ranging from (1, 10), we selected 20 as the number of hidden nodes. In fact, these rules are used as a reference in order to determine the relationship between the number of nodes required in each layer, and the number of hidden nodes selected with different rules will yield avery different value. Our network then has three layers consisting of 369, 20 and 8 nodes.

Next step, text representation to transform the text into a representation suitable for the neural network or categorization algorithms to applied. In our research, documents are represented by the widely used vector-space model (VSM), introduced by Salton, Yangand Wong [23]. In this model, create a space in which both documents and term are represented by binary vector, based on term frequency, and indicates the presence or the absence of a particular term in the document. For each training and testing documents, we created the space vectors corresponding tothe 275 training documents, where each space vector had a dimensionality of 369.

The neural network classifier must be trained before it can be used for text categorization. Training of the neural network classifier is done by the back- propagation learning rule based on supervised learning. In order to train the neural network, a set of training documents and a specification of the pre- defined categories the documents belong to are required.

The neural classifiers are constructed and tested using the MATLAB programming environment from The MathWorks Incorporation. The size of the networks and some parameters used in our experiments are given in table 4.

Table 4. The network size and parameters

#Input Nodes	#Hidden Nodes	#Output Nodes	#Epoch	Learning Rate	Momentum
369	20	8	1000	0.1	0.7

The best back-propagation neural network have 20 neuron nodes in hidden layer and is trained through 1000 epochs with learning rate of 0.1 andmomentum 0.7

5. Experimental Results

The corpus of text messages of collaborative learning are classified into 8 categories. The experiment is conducted in 5 times. The performance of each experiment of training set against the percentage of correct categorization of messages is shown in table 5. The average categorization error of 5 experiments back propagation neural network in this framework is of 0.009806 with a standard deviation of 1.12%.

Table 5. Categorization performance

Experiment (*n*)	Acknowledge	Discuss	Inform	Maintenance	Mediate	Motivate	Request	Task	Error
1	93.3%	80.0%	56.0%	83.3%	0.0%	62.5%	50.0%	61.5%	0.00996
2	93.3%	76.0%	56.0%	83.3%	0.0%	75.0%	65.0%	53.8%	0.00997
3	86.7%	76.0%	64.0%	75.0%	0.0%	50.0%	60.0%	53.8%	0.00937
4	80.0%	84.0%	56.0%	83.3%	100%	50.0%	65.0%	53.8%	0.00981
5	93.3%	68.0%	60.0%	83.3%	100%	62.5%	60.0%	69.2%	0.00992

In order to evaluate a neural network task of collaborative learning skill we first define a contingency matrix representing the possible outcomes of the classification as shown in table 6. Several measures in the information retrieval and machine learning have been defined based on this contingency table. Recall, precision and *F*-measure shown in Eqs. (3-5), are the evaluation measures that have been widely applied for performance evaluation and analysis.

Table 6. Contingency table for binary classification

	Category positif (C+)	Category negative (C-)
Assigned positive (A+)	true positive (*tp*)	false positive (*fp*)
Assigned negative (A-)	false negative (*fn*)	true negative (*tn*)

$$\text{Recall} = \frac{tp}{tp+fn} \qquad (3)$$

$$\text{Precision} = \frac{tp}{tp+fp} \qquad (4)$$

$$\text{F-Measure} = \frac{2\,(Recall.Precision)}{(Recall + Precision)} \qquad (5)$$

In Eqs. (3) and (4), tp (true positive) is the number of documents that are correctly categorized, fp (false positive) is the number of documents that are put into a wrong category, and fn (false negative) is the number of categorized documents that actually belong to no category. In Eq. (5), F is a balanced F-Measure, which is a combined measure of Precision and Recall.

In our research, confusion matrix C, which is *n* x *n* matrix for N-class classifier is used to compute recall, precision and *F*- measure. It is 8 by 8 matrix and an element C[a,b] indicates how many documents in category a have been classified as category b. In ideal classifier all entries will be zeroes except diagonal. Figure 4 shows the result of neural network classifier in the 8 by 8 confusion matrix.

Figure 4. 8 x 8 confusion matrix

The performance of text categorization systems can be evaluated based on their categorization effectiveness. The effectiveness are measured by recall, precision and F-measure. We used the micro-average method to obtain the average value of the precision, recall and F-measure. The experimental results are illustrated in table 7 and figure 5 that display the performance on

8 categories from text of collaborative learning skill in online discussion.

The performance measured by information retrieval standards is classified and sorted according to the eight categories in table 6. Different category numbers of eight categories may cause diverse performance measures. Neural network achieved better results in smaller category classes, e.g. Mediate, than larger ones, e.g. Discuss, Inform and Request. The performance of most categories is satisfactory. The micro average value of overall recall rate, precision rate and F- measure rate is more than 70%.

Table 7. Efficiency measures

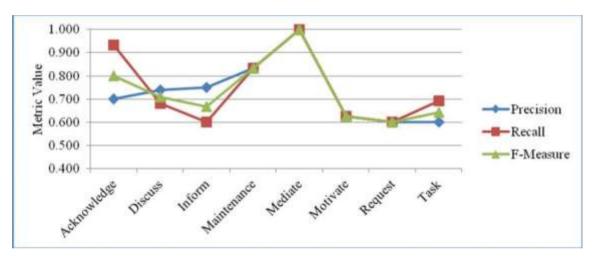| Category	Recall	Precision	F-Measure
Acknowldege	0.933	0.700	0.800
Discuss	0.680	0.739	0.708
Inform	0.600	0.750	0.667
Maintenance	0.833	0.833	0.833
Mediate	1.000	1.000	1.000
Motivate	0.625	0.625	0.625
Request	0.600	0.600	0.600
Task	0.692	0.600	0.643
Microavg.	0.745	0.731	0.735

Figure 5. Evaluating collaborative learning skill with three evaluation measures

6. Conclusion

This paper shows, through experimental results, that neural networks are good classifiers for the text categorization. Experiment is conducted using the proposed framework to categorize the text message of online discussion into eight categories of collaborative learning skill. The average reliability values of categorization messages by human coder is 74.55% while the effectiveness of neural network that are measured by recall, precision and F-measure is also 73.70%. It shows that the accuracy of neural network approach in coded themessages is near with the human accuracy. The results also demonstrate that back-propagation neural network which consist of three layers are able to give good categorization performance.

References

[1] S. Dumais, "Using SVMs for text categorization," in IEEE Intelligent Systems, M. A. Hearst, B. Schölkopf, S. Dumais, E. Osuna, and J. Platt, Eds: Trends and Controversies—Support Vector Machines, vol. 13. 1998

[2] A.K.F. Lui, S.C. Li and S.O. Choy, "An evaluation of automatic text categorization in online discussion analysis," Seventh IEEE International Conference on Advanced Learning Technologies, Niigata, Japan, 205-209. 2007

[3] W. Law and K.F. Low (1997). Automatic document classification based on probabilistic reasoning: Model and performance analysis. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Vol.3, 2719-2723.

[4]   Soucy, P & Mineau, G.W. (2001). A simple k-NN program for text categorization. The first IEEE International Conference on Data Mining (ICDM_01), Vol. 28, 647-648.

[5]   Ma, L, Shepherd, J & Zhang, Y (2003). Enhancing text clasification using synopses extraction. The fourth International Conference on Web Information System Engineering. 115-124.

[6]   Farkas, J. (1994). Generating document clusters using thesauri and neural networks. Canadian conference on Electrical and ComputerEngineering, Vol.2, 710- 713.

[7]   Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. 10th European Conference on Machine Learning, 137-142.

[8]   Benkhalifa, M., Bensaid, A. & Mouradi, A. (1999). Text categorization using the semi-supervised fuzzy c- means algorithm. 18th International conference of the north american fuzzy information processing society-NAFIPS, 561-568.

[9]   Kazama. J & Tsujii, J. (2005). Maximum entropy models with inequality constrains: A case study on text categorization. Machine Learning. 60(1-3), 159- 194.

[10]   Gabrilovich, E. & Markovitch, S. (2004). Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5. Proceedings of the 21st International conference on machine learning, 321- 328.

[11]   Yu, B., ben Xu, Z., hua Li, C. (2008). Latent semantic analysis for text categorization using neural network. Journal of Knowledge-Based Systems, 21, 900-904.

[12]   Ng, H.T., Goh, W.B. & Low, K.L. (1997). Feature selection, perception learning, and usability case study for text categorization. 20th Annual international ACM-SIGIR Conference on research and development in information retrieval, 67-73.

[13]   Savio L. Y. Lam & D. K. Lee (1999). Feature reduction for neural network based text categorization. Sixth International Conference on Database Systems for Advance Applications(DASFAA), 195-203.

[14]   Gokhale, A.A. (1995). Collaborative learning enhance critical thinking. Journal of Technology Education, 7(1).

[15]   Soller, A. (2001). Supporting social interaction in an intelligent collaborative learning system. Journal of Artificial Intelligence in Education, 12.

[16]   McManus, M, & Aiken, R. (1995). Monitoring computer-based problem solving. Journal of Artificial Intelligence in Education, 6(4), 307-336.

[17]   Hidalgo, J.M.G. (2003). Text representation for automatic text categorization, Eleventh conference of the european chapter of the association for computational linguisitics EACL 2003. <http://www.conferences.hu/EACL03/>.

[18]   Porter, M.F. (1997). An algorithm for suffix stripping. In Morgan Kaufmann, Readings in information retrieval, (pp. 313-316). Morgan Kaufmann Publishers Inc.

[19]   Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Survey, 34(1), 1- 47.

[20]   Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Proceedings of ICML- 97, 14th International conference on machine learning, 412-420.

[21]   De Wever, B., Schellens, T., Valcke, M. & Van Keer, H. (2005). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review, Computers & Education 46(1), pp. 6–28.

[22]   Krippendorff (2004). Quantitative content analysis: An introduction to its method, Beverly Hills, Sage Publications.

[23]   Salton, G., Wong, A. & Yang, C.S. (1975). A vector- space model for automatic indexing. In Communications of the ACM, Vol.18, Issue 11 (pp. 613–620). ACM Press.