Next learning topic prediction for learner’s guidance in informal learning environment

Mohammad Sadegh Rezaei*
School of Engineering,
Department of Network Sciences and Technology,
Faculty of New Sciences and Technologies,
University of Tehran,
Tehran, Iran
Email: ms.rezaei@ut.ac.ir
*Corresponding author

Mohammadmehdi Yaraghtalaie
Information Technology Department,
Tarbiat Modares University,
Tehran, Iran
Email: m.talaie@modares.ac.ir

Abstract: Estimating the learning needs of learners in a Social Learning Network (SLN) is very important in proper planning for improving learning space. This paper presented a predictor to estimate the learning needs of learners in SLNs. In Question & Answer Networks, estimating the need for learning means estimating the future subject of the question. The significance of the similarity of the sequence of previous learning subjects with the future subjects of learners is one of the most important areas for estimating the subject of future learning. Hence, this predictor estimates the next learning subject based on the similarity of the subjects about which the learner asks questions. The estimation method introduced in this study is based on the Bayesian solution method. The performance of this method was evaluated in the dataset extracted from one of the most widely used SLNs. The results showed that the proposed method was able to detect future tag of each learner with 78% precision in the informal learning environment using the tags of the questions asked by learners.

Keywords: social learning network; learning topic prediction; Q&A website; online informal learning; interactive learning environment; StackOverflow; technology enhanced learning; community of practice; group forming; learning behaviour; tag prediction.


Biographical notes: Mohammad Sadegh Rezaei received the BS degree in Information Technology Engineering from the Shiraz University of Technology, Shiraz, Iran, in 2011, the MS degree in Information Technology Engineering from Tarbiat Modares University, Tehran, Iran, in 2013, where he is currently working toward the PhD degree from Tehran University. He was awarded Iranian National Elites Foundation scholarship in 2017. Also, he is a
Next learning topic prediction for learner’s guidance

Senior Lecturer in the Information Technology Department at the Shiraz University of Technology and the School of Engineering at the University of Jahrom.

Mohammad Mehdi Yaraghtalaie received the B.S. degree in Information Technology engineering from the Shiraz University of Technology, Shiraz, Iran, in 2013, the M.S. degree in Information Technology engineering from Tarbiat Modares University, Tehran, Iran, in 2016. His research interests are Technology Enhanced Learning (TEL).

1 Introduction

The expansion of the uses of social networks, the development of the level of access to cyberspace, and improvement in the quality of services in this space have provided many opportunities to increase the quality of life of people in the context of cyberspace (Arnaboldi et al., 2017; Brandão and Moro, 2016; Coelho and Duarte, 2016; Maher et al., 2016). One of these social networks is social learning networks (SLNs) that have organised the participation of individuals to meet their learning needs (Badri et al., 2017; De Meo et al., 2017; Vassileva, 2008). In these networks, people chat about various topics and exchange their favourite skills and knowledge in this space. Learners in this space try to reach learning content appropriate and relevant to their learning goal, or by forming learning groups with people who share their learning interest and have some knowledge and skill in that area, they engage in learning and exchange ideas (Vassileva, 2008). A successful platform for SLNs is Q&A networks such as StackOverflow that organise diverse learning environments across a variety of topics, and thousands of people interested in various topics exchange knowledge in this environment, meet their diverse learning needs, and create an informal learning environment (Brinton and Chiang, 2014; Mocker and Spear, 1982; Zheng et al., 2015). One of the major challenges in these networks is directing learners to various resources and groups of learning in the network and helping them create appropriate learning groups consistent with their learning goals and needs (Vassileva, 2008; Zheng et al., 2015). The existence of a million unanswered questions in StackOverflow learning network and the high dropout rates in MOOC learning network indicate the seriousness of this challenge in increasing the efficiency of SLNs and increase of the motivation of learners in this informal learning environment (Brinton and Chiang, 2014; Riverin and Stacey, 2008).

The purpose of this study is to help learners present in SLNs by estimating the future subject of learning in informal learning environments. Predictor of future learning subjects provides the learners with the right context for their automatic guidance in the learning network. On the other hand, given the high rate of drop-off in MOOCs and thousands of unanswered questions in Q&A website, it seems that estimating the learning needs of learners can play an effective role in directing them to learning content, learning colleagues, and appropriate educators to support learning.

The structure of this paper is as follows. The next section will review the research on estimating the needs of learners in SLNs. The third section explains the proposed method to estimate the learning needs. Section 4 evaluates the accuracy of the proposed predictor in SLN and StackOverflow question and answer website and the results are described. In the last section, we will draw conclusions from the discussion and description of future research.
2 Literature review

The research area of estimation in SLNs focuses on estimation of learners’ performance and the estimation of drop-off rate of learners.

2.1 Estimating the performance of learners

Estimating the performance of learners focuses on creating the ability of learning systems in estimating the performance of learners in SLNs through information from evaluations such as homework and quizzes (Brinton and Chiang, 2014). Estimating the performance of learners in the learning environment is done by examining the knowledge needed to perform specific learning activities, the results of other learners’ performance in learning activities, and learners’ experiences and views about learning activities (Thai-Nghe et al., 2010). The easiest way to estimate learner’s performance is to provide a classification model that classifies learners based on effective performance parameters and estimates the grades of other learners who are in the same group based on the grades of some learners in learning activities. For this purpose, factor analysis techniques are used to extract the relations between the question-concept and the users’ skill to determine the relations of questions and the efficiency in skills and activities. Traditional class data is used to perform factor analysis. In studies by Bergner et al. (2012) and Lan et al. (2014), factor analysis was used to examine the relation between the behaviour of learners in the network and their efficiency. These two studies have tried to estimate based on the relations found and the performance of learners in the network based on their observed behaviours in the evaluations. Research by Bergner et al. (2012) has used collaborative filtering methodology to estimate performance. Research by Lan et al. (2014) has proposed a method based on machine learning to estimate the knowledge of learners in a specific field. The proposed model in this study calculates the probability of the learner’s correct answer to the questions and estimates the learner’s performance in the learning environment. This research has used one of the Maximum Likelihood Estimation methods and Bayesian solution for factor analysis and finding the factors of estimation. The most important challenge in estimating efficiency is the lack of sufficient information about the activities and evaluations of learners that creates the problem of “cold start” (Brinton and Chiang, 2014; Thai-Nghe et al., 2010). This problem is much more common in SLNs created by MOOCs. Although there are plenty of user evaluation data available in MOOCs, overall, the number of evaluation questions each learner answers is very limited. For example, in Networks: Friends, Money, and Bytes (N: FMB) period where MOOC is presented, only 9% of learners respond to at least one question. Hence, the use of the results of the answer to the questions for estimation will not yield a precise result, although the use of very rich social data in MOOCs can greatly increase the accuracy of estimations (Brinton and Chiang, 2014).

2.2 Estimating drop-off rate of learners

Research conducted in this section deals with estimation of the rate of drop-off of learners in online courses in order to improve the course and retaining of learners in the learning environment. Calculating drop-off rate can be calculated for each learner or the sum of all students in the course (Cheng et al., 2013). SLNs data can be used to determine drop-off rate among learners. Therefore, the research trend in this area is to use
Next learning topic prediction for learner’s guidance

65

this information to estimate individual and group drop-off rate of learners in SLNs. By observing activity reduction in participation rate of a person who is active in terms of question and answer questions in the learning environment, one can estimate that he will not fully do this activity. In addition, for the learners who are not active in the polls, it is possible to estimate their future by continuing or abandoning their activities using similar learners in the network (Brinton and Chiang, 2014). Cheng et al. (2013) have introduced a tool to estimate the risk of drop-off of learners in large learning environments. By providing an estimation of the probability of drop-off of learning, the tool provides the ability to provide appropriate learning support to maintain learner motivation.

SLNs, such as question and answer websites, have been introduced as informal learning environments that can be developed and strengthened by the communities of the operand in pursuit of the learner’s learning objectives. Estimating learning topics in these learning environments can create the right context for creating and strengthening independent and diverse learning groups in these networks and increase the efficiency of these networks in achieving effective learning.

3 The explanation of the proposed method

The proposed method for estimating future learning is based on Bayes’ probability theorem. In this method, based on the tags of questions learners have asked and calculating the frequency of their similarity, the probability of occurrence of questions with probable tags is measured and based on that future tag is estimated. The learner is expected to ask questions in the area of estimated tags in the future. The steps for estimating the proposed method are described below.

In the first step, first, the questions that learners have asked are specified separately for each learner, and then the tags related to the questions assigned to each learner are extracted. After that, the repetitive tags are deleted.

In the second step, we identify users similar to the ones whose future topic of learning we want to estimate (for more accurate guidance) in the network, from the similarity of the existence of the tags related to the questions asked. Subsequently, the tags selected by similar learners but not by the intended learner are considered as candidate tags. Candidate tags are future questions the learner is likely willing to participate in.

In step 3, the probability of occurrence of each candidate tag is calculated based on the number of similar learners with that tag, and the tag with the highest probability is considered as the learner’s future tag. For this purpose, the learner’s future tag is calculated from equation (1).

\[
p_{j}(\text{tag}_j) = \frac{\text{sim}_j(\text{tag}_j)}{|\text{sim}_j|}
\]

where \(p_{j}(\text{tag}_j)\) is the probability of interest and knowledge need for a question with \(\text{tag}_j\) by the \(i\)-th learner. \(\text{sim}_j\) is the set of learners who have asked questions covering the questions asked by \(j\)-th learner. In other words, similar to this learner, they are a part of their learning path. \(|\text{sim}_j|\) is the number of members in this set. \(\text{sim}_j(\text{tag}_j)\) is the set...
of learners similar to the $j$-th learner (members of $sim_j$ set) who have asked questions with tag $tag_j$.

If $r'$ is the set of tags related to the questions asked by $j$-th learner in the network and matrix $x$ is the tags related to the questions asked by all the network learners, so the row $x_z$ is the matrix of the set of questions asked by the $z$-th learner ($x_{z,k}$ shows $k$-th tag from learner $z$-th). Moreover, $y^k$ is the matrix of similarity of the $j$-th learner to $k$-th learner in the tags. If the tags corresponding with input $x_i$ are in the set $r'$, its corresponding entry in matrix $y^k$ is equal to 1 and otherwise equal to zero. If the internal multiplication of the matrix $y^k$ and its transpose ($(y^k)'$) are equal to identity matrix ($I$), the two learners are similar in the tags (the questions asked by the $k$-th learner include all the tags of the questions asked by the $j$-th learner). In fact, determining the members of the set $sim_j$ is calculated by equation (2).

$$\left(y^k\right)' * y^k = I \Rightarrow l_i \in sim_j$$

where $l_i$ is the vector of the tags corresponding to the questions asked by the $k$-th learner. The set $T'$ contains all $l_i \in sim_j$ tags (sets of the tags of the learners similar to $k$-th learner) that do not exist in $l_j$ and will exist as candidate tags that can exist in the set of $j$-th learner’s tag. $sim_j\{tag_i\}$ is a set of learners who are members of $sim_j$ set and have the tag $tag_i \in T'$ in their tag set.

### 3.1 Improved version of the proposed method

In addition to tags, the number of replicas of each tag in the learner’s knowledge set is also a good indicator of how much interest and effort the learner has in different subjects. Therefore, in the other version of the proposed method, the number of iterations of tags is considered in the next tag estimation. Accordingly, learners whose number of iterations of tags is more similar (indicating the number of questions followed by the learner with the related tag) are more likely to be more influential in estimating future tags. Therefore, an index called $sim_{R_{kj}}$ is defined as the degree of similarity of $j$-th learner with $k$-th learner in iteration of similar labels. This index calculates the similarity value in proportion to the maximum difference in the intended tags in accordance with equation (3).

$$sim_{R_{kj}} = \sum_{i \in l_j} D_{kj}(tag_i) \text{ tag}_{i} \in L_j$$

where $D_{kj}(tag)$ is the absolute value of the number of iterations of tag that are in the tag set of $j$-th and $k$-th learners, which all represent the difference between iterations of common tags between $j$-th and $k$-th learners. Therefore, the probability of the occurrence of $i$-th tag from the set of candidate tags is calculated from equation (4).

$$p_j\{tag_{i}\} = \frac{\sum_{k=1}^{N} sim_{R_{kj}}}{\sum_{r \neq t}^{N} sim_{R_{rj}}} \text{ if } l_i \in sim_j\{tag_{i}\} \text{ and } l_p \in sim_j$$

The set $T_j$ contains all $kjlsim \text{ set}$ tags (sets of the tags of the learners similar to $k$-th learner) that do not exist in $lj$ and will exist as candidate tags that can exist in the set of $j$-th learner’s tag.
This version of the proposed method is expected to be more accurate in estimating the subject of future learning.

4 Evaluation and results

StackOverflow Question and Answer website has been used to evaluate the proposed method. This website was launched in 2008 and provided a network of learners in the field of computer programming. The website now (in 2017) has more than 7 million users, in which 14 million questions have been asked and over 22 million responses have been registered (see https://stackexchange.com/sites#). Considering the extent and popularity of this website among programmers of this environment is considered as one of the most effective learning environments for informal learning (Zheng et al., 2015). A dataset from this website has been published by the University of Calvia Irvine School of Computer Science, which is provided from the information of this website from 18 February 2009 to 8 June 2009. For many years, this dataset has been a reference to evaluate the methods presented to estimate the right answers to questions, to estimate the period until the full answer to the question, and to find the experts (DuBois, n.d.). This dataset contains 8343 questions raised, 26,837 users asking the questions, and 26,752 respondent users, including 26,354 records. The data in this dataset is shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Datasets analytical information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of learners</td>
<td>37,255</td>
</tr>
<tr>
<td>Number of questioners</td>
<td>26,837</td>
</tr>
<tr>
<td>Number of answers</td>
<td>26,752</td>
</tr>
<tr>
<td>Number of questions</td>
<td>83,423</td>
</tr>
<tr>
<td>Number of answers</td>
<td>263,540</td>
</tr>
<tr>
<td>Average of number of question asked by each learners</td>
<td>3</td>
</tr>
<tr>
<td>Average of number of answer by each learners</td>
<td>8.9</td>
</tr>
<tr>
<td>Average of answers to each questions</td>
<td>16.3</td>
</tr>
<tr>
<td>Average of questioners to each question’s tags</td>
<td>5.10</td>
</tr>
<tr>
<td>Average of answerers to each question’s tags</td>
<td>2.26</td>
</tr>
</tbody>
</table>

This dataset includes the ID of the learners and responders to questions, the ID of the questions raised, identifier, the ID of the answers provided to the questions, and the ID of the topics to which the questions relate and the time of the question or answer to the questions.

In order to assess the proposed method from among the learners who asked at least 10 questions, 20% of them (approximately 1400 learners) were randomly selected. The tag of the last question asked by these learners was considered as the tag that the proposed method should estimate.

The accuracy of the two versions of the proposed method in estimating the future tags of learners is shown in Figure 1.
As is seen, in a version of the proposed method that does not consider the iterations of the tags for measuring similarity, about 65% of the tags are correctly estimated, which is 78% in the version that considers the iteration of the tags. These results indicate that due to the various patterns of questions raised, the estimation of future issues largely depends on the number of questions asked by other learners and the number of iterations. This is thus because many learners are present on this website and due to the unofficial nature of this learning environment, they have various learning behaviours. The remarkable point is that considering the number of iteration of labels (which indicates the depth of a learner’s focus on a particular subject) improves the accuracy of the estimations. This fact shows that future subjects of learners’ are dependent on the frequency of topics, in addition to the nature of previous topics.

About 2% of the tags in the version not considering the iterations did not obtain the highest probability to fit among the correctly estimated tags, but they were among the candidate’s tags for estimating and obtained less probability. This number reaches 3% in the version that considers the iterations. It seems to be possible to estimate these tags accurately by increasing the accuracy of the method and the use of social information in the network.

About 33% of the tags in this version not considering the iteration of the tags, and about 20% in the version considering the iteration of tags are not correctly estimated. The tags of this number of learners were not even present among the candidate labels. Two reasons for this fact are imaginable. First, the proposed method is unable to identify tags that are likely to occur, and the future tags of these learners are dependent on information other than the tags of the set of question asked. Secondly, these learners pursued a new area and a new model in the network that had nothing similar to in the network, which, given the informal learning environment and the freedom of learners in determining their own learning path based on their interests and needs, it is not unimaginable. Accordingly, the future tags of such learners cannot be estimated based on the information available on the network, so the lack of proper estimation of the future subject of these learners is not a weakness of the proposed method.
5 Conclusions and future work

The results of this study indicated that the similarity of learners’ knowledge sets is very effective in future tags. Hence, the proposed predictor, designed based on Bayes’ Theorem, has been able to estimate the future tags with a precision of 78% for a wide variety of learning logs. However, it should be noted that some learning topics in the informal learning environment are very new and not similar to topics that learners have followed in the past thus unpredictable. Given this, the accuracy of the estimation is very good. Moreover, the proposed method, in addition to a variety of topics, considers the depth of the attention of learners for measuring similarity. Using the results of this study, one can provide services to guide learners in the network based on their behavioural similarity in asking question and reduce their meta-cognitive barriers in achieving higher performance in the learning network environment based on the experiences of other learners.

Learning areas that consist of two or more learning topics have a great variety in SLN, and given the diversity of learners’ needs and interests, pursuing these learning areas by learners will vary. Hence, the estimation will be more accurate within the learning spheres as the sequence of topics in each learning area is necessary to understand the learning objective in that area. In this research, the predictor estimates the future subjects of learning based on the entire sequence, regardless of the order in which they are arranged. If, given the meaningful relations of the sequence of topics to each other, the similarity of the behaviour and interests of learners, who have the same background in pursuit of topics, considering the sequence of issues the learner has asked about the subject or the answers to the question of the subject and voted for it, can be a good indicator of the future issues that the learner faces. Therefore, designing a predictor that utilises this in estimating feature learning is considered to be of the future studies.

References


DuBois, C. (n.d.) *StackOverflow Data*, University of Calvia Irvine School of Computer Science.


