**Education and Information Technologies**
https://doi.org/10.1007/s10639-020-10155-x

# Adaptive Security Awareness Training Using Linked Open Data Datasets

**Zheyu Tan · Razvan Beuran · Shinobu Hasegawa · Weiwei Jiang · Min Zhao · Yasuo Tan**

Received: 2 November 2019 / Accepted: 12 March 2020

**Abstract** Cybersecurity is no longer an issue discussed only between the professionals or technologists, but it is also closely related to ordinary people whose daily life is exposed to kinds of cyberattacks. And Womabat Security Technologies conducted a survey revealed that ransomware is an unknown concept to nearly two-thirds of employees. In practical, almost 95% of cybersecurity attacks are due to human error. At fact, expensive and sophisticated systems cannot work effectively without considering the human factor, while human factor is the major vulnerability in cybersecurity. Thus, it has great significance to give people cybersecurity awareness training. In this paper, we present a system, named ASURA, providing adaptive training aimed at improving cybersecurity awareness of people. Three issues can't be neglected in adaptive cybersecurity awareness training, as follows. Firstly, we need to decide the proper training contents from the huge training materials. Secondly, the training contents should be timely updated, as cyber attacks constantly changing. At last, we should conduct training through effective and acceptable approach. We solved above three issues in this paper, and the innovative idea of this paper is constructing hierarchical concept map from the LOD database DBpedia. Then, we employ a series of processing on hierarchical concept map, including PageRank algorithm used to calculate the importance of each concept node, and filtering used to filtered out undefined and unrelated concepts. In particular, we get training contents from DBpedia dynamically and timely updated, so that training contents is keeping up to date. ASURA delivered training contents completely online, thus significant trimmed budget and allowed learners accessing training outside of a traditional classroom. Moreover,

Zheyu Tan
Japan Advanced Institute of Science and Technology, Nomi, Japan. E-mail: jotan@jaist.ac.jp
Razvan Beuran, Shinobu Hasegawa, Min Zhao, Yasuo Tan
Japan Advanced Institute of Science and Technology, Nomi, Japan
Weiwei Jiang
School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia

ASURA provide adaptive training targeted to individual learner, as it generate training contents based on the keyword from the learner.

**Keywords** Cybersecurity education · Hierarchical concept map · Relevance estimation · Concept filtering · Adaptive training

## 1 Introduction

Along with the rapid development of technologies, our daily life is exposed to kinds of cyber-attacks, such as password attacks, malware attacks, phishing emails and, etc. In a survey carried out in 2016 by MediaPRO revealed that 88% US employees lack the awareness to prevent common cyber incidents [1]. A later report in 2018 revealed an even worse awareness state. Specifically, surveyed employees did significantly worse in identifying malware warning signs, knowing how to spot a phishing email, and practicing social-media safety [2]. Furthermore Wombat Security Technologies conducted another survey showed that almost a third of employees are lack of knowledge of phishing, so of ransomware for two-thirds of those employees [3]. More importantly, FraudWatch also pointed out that almost 95% of cybersecurity attacks were due to human errors and as much as 54% of companies have experienced one or more cyber attacks in the last 12 months. The number rises every month [4]. Therefore, it is severe that we are confronting with unprecedented challenges in cybersecurity.

For alleviating this issue, a considerable amount of money spent on cybersecurity to protect people, companies and organizations from cyber attacks. However, expensive and sophisticated systems cannot work effectively without considering the human factor, while human factor is the major vulnerability in cybersecurity [5]. Thus it is urgent to improve people's cybersecurity awareness and train those people to understand basic cybersecurity concepts and necessary policies [6].

There are a lot current cybersecurity awareness training programs provided. Especially, some large companies provide regular cybersecurity awareness training for their employees annually or biannually. Currently, there are multiple types of cybersecurity awareness training approaches [7] [8]. For instance, the breakroom approach is widely adopted, in which people are gathered at the break time and are told basic tips about cybersecurity. Alternatively, the security video approach is also popular which shows short cyber security training related videos to people, along with the simple course approach which is made up of simple pdf files of DON'Ts and DOs. Nevertheless, existing systems are confined to a set of matters. Fo instance, the limited training contents, outdated topics, tight budget, location constraints, training efficiency, and let alone personalized targeted training.

For addressing this issue, in this paper, we propose ASURA (Adaptive SecUrity awaReness trAsining), an adaptive cybersecurity awareness training system in cooperated with Linked Open Data(LOD) datasets [9]. ASURA system can query and construct Computer security hierarchical concept map from

the LOD database DBpedia dynamically and timely. In this sense, training contents are up to date and have a considerably wide scope. ASURA delivered training contents completely online, thus significant trimmed budget and allowed learners accessing training outside of a traditional classroom. Moreover, it has been shown that adaptive training is one of the most effective methods to improve cybersecurity awareness of people and reduce potential cybersecurity attacks in daily life [10]. Furthermore, ASURA provide training targeted to individual learner, as it generate training contents based on the keyword from the learner.

The Computer security hierarchical concept map is built form the LOD datasets DBpedia which is one of the most meaningful and famous Semantic Web projects and the most popular open knowledge dataset [11]. DBpedia is constructed by extracting data from Wikipedia which is available in 125 languages, and together describe 38.3 million things that covers many domains. Consequently, DBpedia has an extensive topic coverage. It can further enable people to get much further training materials as it also interlinkes with other kinds of open datasets [12] [13]. For accessing the data in DBpedia which is in RDF graph, a public SPARQL endpoint is adopted [14]. In this paper, we use SPARQL to query information from DBpedia [15]. In particular, the relationships and the properties of information are retrieved and returned in JSON format to build the big hierarchical concept map. We will also discuss about the querying strategy in order to build an efficient and useful concept map for adaptive cybersecurity awareness training.

Once the hierarchical concept map has been constructed, we then process it for the later practice training. We construct the hierarchy concept map from the DBpedia Categories [16]. However, the subconcept map generated from the Computer security map can still be large. For instance, the Malware subconcept map contains 54 concept nodes which may not be significant or appropriate for the training. Therefore, we need to decide which concepts have higher priority for Malware adaptive training for improving Malware knowledge of people. Here, we estimate the concept importance/relevance by applying PageRank algorithm on the Computer security concept map. The PageRank algorithm can be used to determine the relevance or importance of a page, and the importance of a page is determined by the number of links going out of this page.

Finally, we employ filtering on concept map to exclude those nodes with undefined concepts in DBpedia, or nodes which are not so related to the training keyword. The processed concept map is combined with the simple learner model [17] provided the idea of the adaptive training [18]. Question creation and text processing are then performed in training system into actual practice.

In summary, the main contribution of the present paper are as follows:

1. Propose a method to build a hierarchical concept map timely updated from the Linked Open Data (LOD) database DBpedia and extracted subconcept map from it for adaptive training.

2. Propose methods to process the built hierarchical concept map, by employing the PageRank algorithm to calculate the importance of each concept node and filter algorithm to filter the irrelevant or no definition concept nodes on the concept map. The concept map is used for adaptive awareness training later.
3. Propose a simple method to conduct adaptive training and implement an adaptive awareness training system prototype, ASURA. The processed concept map is combined with the simple learner model to provide the idea of the adaptive training.

The rest parts of this paper are organized as follows. In Section 2 we discuss the overall requirements and demonstrated the brief design of ASURA. Then in Section 3 and 4, we present details about the construction and the processing of the concept map. Next, in Section 5 and 6, we discuss training details, and evaluate the research work from several aspects later in 7. This paper continues with conclusions, and ends with acknowledgements and references.

## 2 Motivation and overview

In this section, we describe the main characteristics of cybersecurity awareness training, and present an overview of our approach for addressing the identified issues.

### 2.1 Adaptive cybersecurity awareness training

The objective of adaptive cybersecurity awareness training is improving security awareness of people. Thus those financial loss or information damages caused by the human factors can be ignored. Considering the fact that there are abundant training topics but limited time and energies, the need to decide what concepts should be given to people to do the awareness training comes at the very first. Also, repetitive and adaptive training is also an important characteristic, which makes the training program more effective in practice.

With these insights, the primary motivation of our research is to develop a system that can provide cybersecurity awareness training. The system tends to create change in knowledge of learner so that bring about an overall change and promotion regard to effective cybersecurity management. In consequence, we designed the system,ASURA, so that it can obtain training contents form DBpedia timely updated, and also can decide the training priority of concepts, as well as conduct adaptive training.

In addition, four components are required for the adaptive training as following[19].

1. The expert model contains training materials that used to teach students. The materials can be questions and solutions, or lessons and tutorials.

2. The learner model (referred to as student model) contains information of the learners, such as domain knowledge and learning performance. The learner model can be used to improve practice training.
3. The instructive model combines the previous two models to show the following training content.
4. The instructional environment is a user interface or training platform.

In a training project, the training contents is always a critical key that leads the training to be useful and effective [20]. For the cybersecurity awareness training, getting updated training topics is extremely important due to the rapid change of information technologies. Novel hacking methods or cyber-attacks may bring up destructive damage to our daily life at any moment. Maintaining training materials up to date is one of the critical requirements to conduct practical training to learners.

More crucially, giving learners a brief Malware awareness instruction is not easy, considering the fact that there are ample and plentiful concepts related to Malware. Also teaching all related concepts are not feasible, with limited time and energy budget. Consequently, selecting training concepts with high priority for training is another crucial requirement regarding a resultful training process.

Finally, learners that participate awareness training are different individuals with various education background and security knowledge level. Thus providing learner targeted training is also a vital requirement related to practical training. Providing adequate interaction with learners, and generate training contents depending on the will of learners may also increase the training effect.

In conclusion, the following three considerations are necessary for adaptive cybersecurity awareness training:

1. Getting timely updated training contents.
2. Selecting training concepts with high priority for training considering the limited resources and time.
3. Providing fully interaction with learners, and generating training contents with respect to the will of learners, which personalizes each training for individual learners.

### 2.2 Our approach

We developed ASURA based on the above consideration for the adaptive cybersecurity awareness training. For fulfilling Consideration #1, we propose a method to construct a Computer security concept map from the DBpedia timely updated. For Consideration #2, we employ PageRank algorithm on the above concept map to calculate the importance of each concept. Finally, for the Consideration #3, we generate training content and deliver those contents by the method of adaptive training.

An overview of ASURA is shown in Fig. 1, and the detailed system design of ASURA and the arrangement of this paper is described in Fig. 2. In the

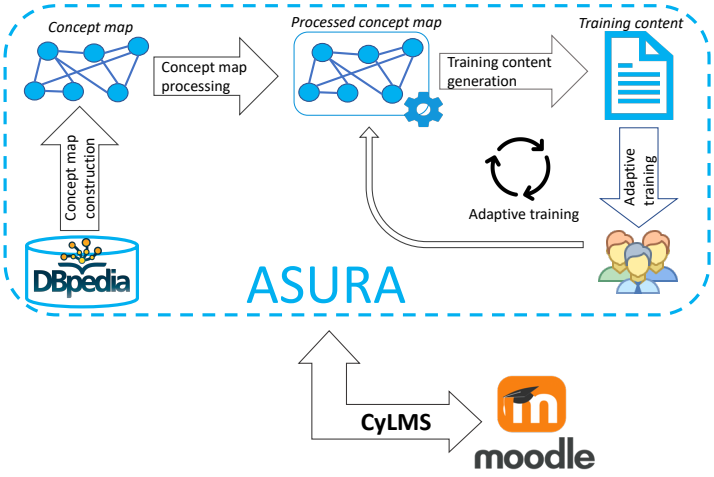


**Fig. 1** An overview of ASURA, and interaction with Moodle

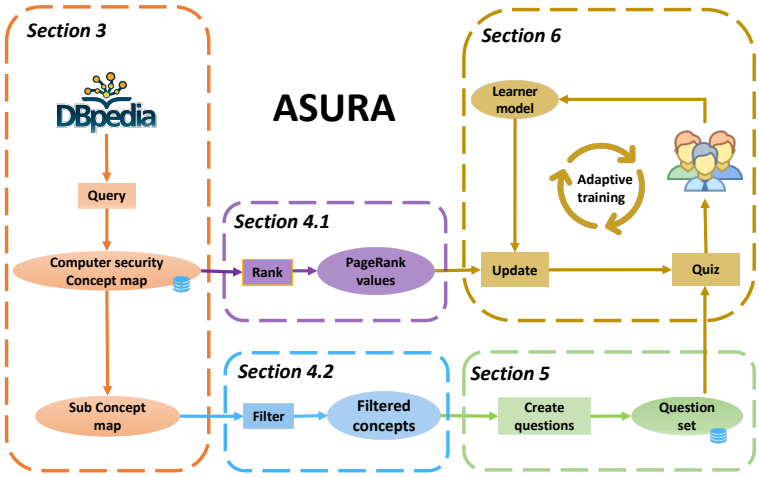


**Fig. 2** System design of ASURA

first stage, we construct a concept map from the LOD dataset DBpedia. This process can dynamically gets much related and timely updated cybersecurity training concepts. The detailed description is described in Sect. 3. In the second stage, we conduct a series of processing to use the concept map for training, with more details discussed in Sect. 4. In the next stage, we generate exact questions for adaptive training in Sect. 5 as clarified. Finally, in Sect. 6, we describe the mechanism of adaptive training, as well as the learner model and the update algorithm. More details about interaction with CyLMS [21] are also included, for which we take training content from database and import to the Moodle LMS.

## 3 Hierarchical concept map construction

In this section, we explain the method to build the hierarchical concept map from DBpedia and discuss the query algorithm applied for ASURA.

### 3.1 DBpedia Query via SPARQL

Conventionally, most websites are suffering from the limited access to valuable but unstructured data, such as Wikipedia. In contrast, DBpedia that convert Wikipedia contents to Resource Description Framework (RDF) [22] enables sophisticated queries against DBpedia datasets derived from Wikipedia [14]. SPARQL is a DBpedia query language, which can acquire the relationships and the properties of structured data. The structured data can be used for the further processing, for which we chose the JSON type from different available formats [23], for its generality and versatility.

In particular, in this paper, for generalizing the hierarchical architectural concept map, we choose the DBpedia Categories dataset . The DBpedia Categories dataset includes information which concept is a category and the relationship between each two categories [14]. The relationship used in our SPARQL queries is: *A skos : broader B*, an essential hierarchical relationships between two concepts [24]. In this relationship, B has a more general or broader meaning than A. Broader concepts are typically rendered as parents in a concept hierarchy. For instance, *Computer_security skos : broader Security* means that concept Security has a broader meaning than concept Computer_security, thus concept Security is considered as a parent. The core query statement used in this paper is as follows:

```
SELECT DISTINCT ?child ?childlabel
WHERE{
?child skos:broader ⟨http://dbpedia.org/resource/Category:concept⟩;
rdfs:label ?childname.
FILTER (LANG(?childname) = 'en')
BIND (?childname AS ?childlabel)
}
```

### 3.2 Query Strategy

Using before aforementioned method, we build the entire hierarchical concept map based on a seed keyword: Computer security. The concept map is a collection of entities called nodes representing concepts. The concepts are then connected by edges, which is the property $skos : broader$. In computer science, there are two most common tree traversal algorithms: Depth-first search (DFS) and Breadth-first search (BFS) [25] [26]. We choose BFS search to reach the descendent concepts of Computer security in level order, considering the character of awareness training that does not require in-depth knowing or understanding.

In particular, ASURA finally adopt the seven depths Computer security hierarchical concept map that included 2640 concepts. Moreover, in order to display the concept map visually and directly, Fig. 3 illustrates the two depths Computer security concept map. In summary, this map contains 126 concept nodes in total, in which the root is Computer security, and it has 22 children concepts, with another 103 grandchildren concepts.

Furthermore, it can be observed that with depth growing, the number of nodes/concepts increase rapidly. Consequently, the time for accessing the Internet grow immediately. In particular, with relatively large depth, the time for accessing the Internet is almost hundred times slower than the execution time. More crucially, since the descendant concepts of Computer security is enormous, as shown in Fig. 4 with seven different depths, the number of nodes/concepts increase more significantly.

For address this issue, in ASURA, we update that seven depths Computer security hierarchical concept map periodically, and extract the needed subconcept map from it in practical training. Hence, the seven depths Computer security concept map can be kept up-to-date efficiently and a significant part of time can be saved for training in practice.

## 4 Concept map processing

After constructing the hierarchical concept map, we conduct a series of processing methods to further organization of the contents. The processing includes two parts: concept relevance estimation and filtering, as we describe details in this section.

### 4.1 Concept relevance estimation

Using the SPARQL query, we constructed the hierarchical concept map from the DBpedia Categories and extracted the needed subconcept map from the big one for practical training. However, the subconcept map generated from the Computer security map can still be large. For example, the Malware subconcept map still contains 54 concept nodes which can not be all used for the later

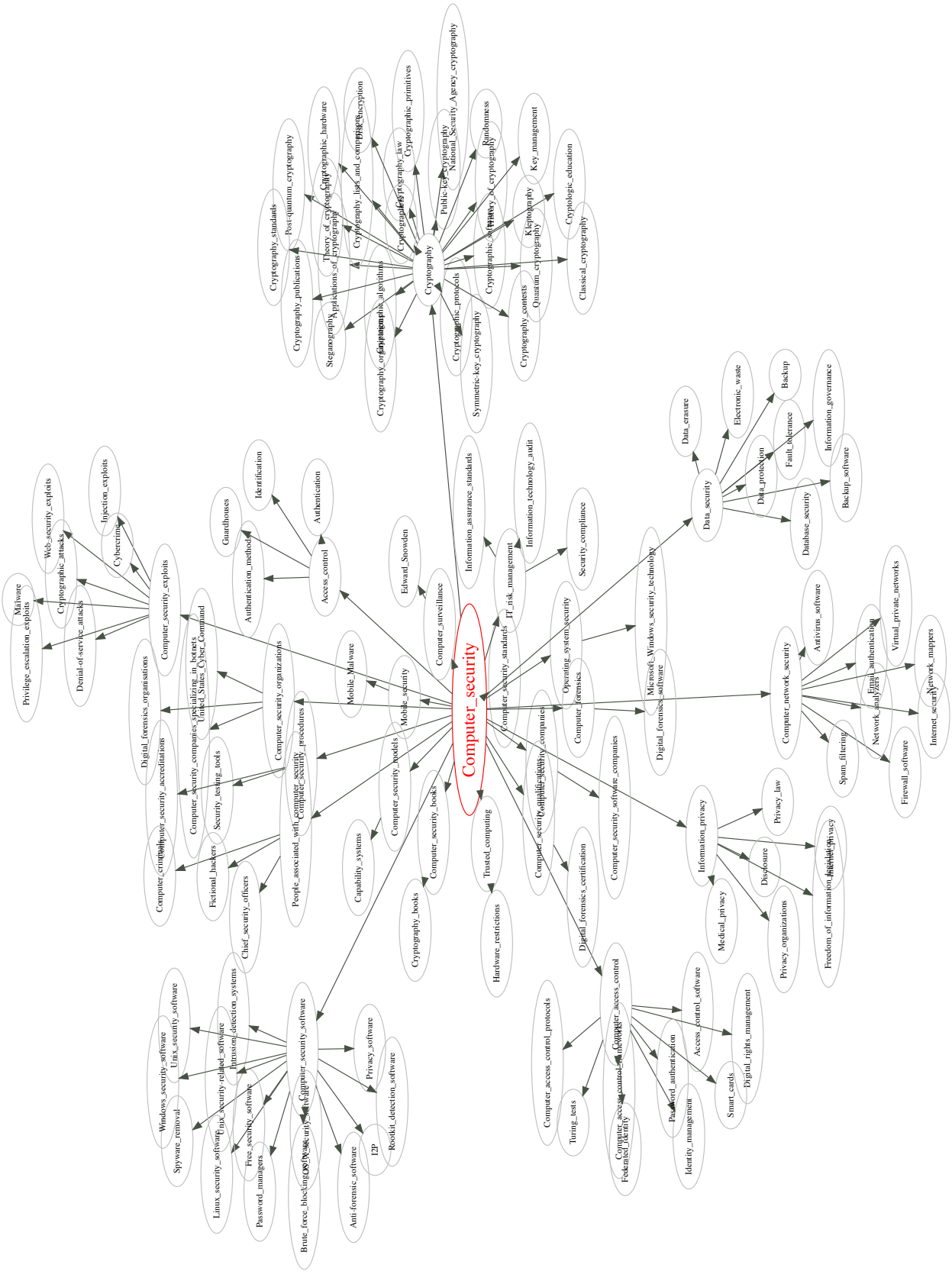


**Fig. 3** Computer security concept map, level = 3

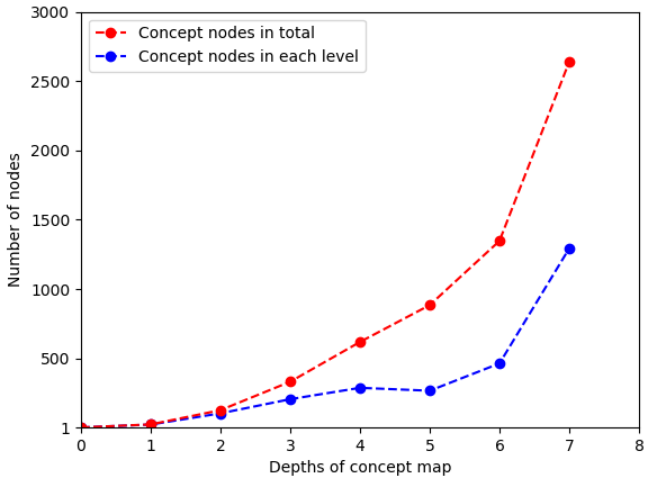


**Fig. 4** Computer security concept map, with various depths

training, considering the fact of limited energy or time of the learner. In this consequence, we need to decide which concepts are more critical or related to the Malware adaptive training. Hence, before using this concept map for adaptive training, the first step is to estimates the concept relevance/importance.

In this paper, we define the importance of a concept node as the number of linked concept nodes going out of this concept node, i.e., the out order of the node. For calculating the importance, we adopt the PageRank algorithm [27] on the concept nodes [28]. By substituting the pages by concept nodes, we can then rank the importance of the concept nodes effectively. In PageRank algorithm, the ranking of a node is recursively given by the ranking of those nodes which linked to it, and the PageRank algorithm is as follows:

*PageRank algorithm* We assume node A has nodes T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. Also, C(A) is defined as the number of links going out of node A. The PageRank of node A is given as follows [27]:

$$PR(A) = \frac{(1-d)}{N} + d(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)}) \quad (1)$$

However, the final PageRank values of those concepts are unknown since the PageRank of a concept is always given recursively by the PageRank of related concepts. This issue can be resolved by the method provided by Lawrence Page and Sergey Brin's paper [27]. Therefore, with a randomly assigned PagePank value of each concept at first, we repeat calculations many times until the values stop changing significantly as $(P_{n+1} < -P_{n+1}) < \epsilon$. When the PageRank

**Table 1** Final PageRanks, seed keyword: Computer security, iterations: 56

| **Concept** | **PageRank** |
|---|---|
| Computer_security | 0.10382910 |
| Access_control | 0.05869410 |
| Cryptography | 0.02845521 |
| Computer_security_exploits | 0.00679187 |
| Data_security | 0.00668850 |
| Information_privacy | 0.00642375 |
| Computer_access_control | 0.00354294 |
| Computer_network_security | 0.00240053 |
| People_associated with computer security | 0.00208197 |
| Computer_security software | 0.00187915 |
| Trusted_computing | 0.00183606 |
| Computer_security_organizations | 0.00042248 |
| It_risk_management | 0.00042248 |
| Computer_security_procedures | 0.00032132 |
| Computer_security_companies | 0.00030615 |
| Computer_security_companies | 0.00030615 |
| Computer_surveillance | 0.00030615 |
| Operating_system_security | 0.00022017 |
| Mobile_security | 0.00022017 |
| Computer_forensics | 0.00022017 |
| Computer_security_books | 0.00022017 |

values converged or reached at a fixed point, the normalized probability distribution will be 1.0. According to Lawrence Page and Sergey Brin, around 100 iterations are necessary needed to get a good approximation of the PageRank values of the whole map.

As an example in Tab. 1, we show part of the final calculated importance/PageRank value on Computer security map after 56 iterations in descending order. The damping factor is set to 0.85, the maximum number of iteration is set to 100 and the $\epsilon$ is set to $10^{-9}$.

### 4.2 Filtering

We then perform a filtering process on the concept map to remove those undefined and unrelated concepts. The process is based on the two factors: concept utility, and the DBpdia class and property.

*Filtering concepts based on concepts utility:* Firstly, we need to remove those undefined concepts based on the hypothesis of that those concepts without definition are not essential for practical training. Thus, we discard no definition concepts in ASURA. In particular, ASURA choose the DBpedia Articles dataset for retrieving concept definitions. The DBpedia Articles dataset includes description of all 1.95 million concepts within the English Wikipedia including titles, short abstracts, thumbnails and links to the corresponding articles [14], from which we can query concept definition. However, in the course

of the definition retrieving, we should not ignore a fact that the spelling of a concept in DBpedia Articles dataset and the DBpedia Categories dataset can be different. For instance, the concept Computer_worms in Categories dataset is different from the Computer_worm in Articles dataset. If we use Computer_worms as the keyword to retrieve the definition, SPARQL engine returns none. In this case, we filtered out computer_worms concept that actually has a definition in DBpedia. To address this problem, we should first link the concept from Categories dataset to Articles dataset, and then retrieving definition. In particular, the linking from Articles to Categories is represented as *dct* : *subject* [29], and the core query statement used in this paper is as follows:

```
SELECT DISTINCT ?child_defi ?childlabel_defi
WHERE{
?child_defi dct:subject ⟨http://dbpedia.org/resource/Category:concept⟩ .
FILTER (LANG(?childname_defi) = 'en')
BIND (?childname_defi AS ?childlabel_defi) }
```

However, the above statement may return multiple corresponding concepts from Articles dataset. For example, the concept Malware in Categories dataset has dozens of corresponding concepts from Articles dataset. Thus, we use a simple Nature Language Processing library -fuzzywuzzy [30] to find the most approximate string matching of the concept among the return results of *dct* : *subject*.

*Filtering concepts based on DBpedia class and property:* Secondly, we need to remove those irrelevant concepts based on the hypothesis of that those unrelated concepts are not essential for practical training. For instance, concept Digimon is a concept from Malware subconcept map, and to be more specifically, Digimon is grandchildren concept of Malware. However, concept Digimon is an instance of Game class, and it is irrelevant to the cybersecurity awareness training. In summary, we discard those concepts in ASURA based on DBpedia class and property. Tab. 2 is the list of irrelevant classes used in this paper.

After performing the filtering process on the concept map, we removed both undefined and irrelevant concepts. For example, Tab. 3 showed the finally filtering results of Malware subconcept map. In summary, 20 related and defined concepts left in Malware subconcept map, as *level* 1 concepts are all filtered out because of no definition in DBpdia Articles, and concepts like Digimon are also filtered out since it is irrelevant to the Malware training.

## 5 Training contents generation

After performing a series of processing on the hierarchical concept map, this section ables to generate the specific training contents with the processed concept map.

**Table 2** Irrelevant classes list

| Concept |
|---|
| movie |
| Movie CW |
| televisionshow |
| animal |
| grape |
| place |
| Planet |
| Location |
| Agent |
| Q386724 |
| game |

**Table 3** Filtered concepts related to Malware

| Concept | PageRank | Level |
|---|---|---|
| Malware | 0.00309623 | 0 |
| Computer_virus | 0.00050839 | 2 |
| Winwebsec | 0.00029768 | 2 |
| Computer_worm | 0.00023484 | 2 |
| AIDS_(Trojan_horse) | 0.00023484 | 2 |
| Linux_malware | 0.00016091 | 2 |
| Dendroid_(malware) | 8.69785912e-05 | 2 |
| Blackhole_exploit_kit | 8.69785912e-05 | 2 |
| Rogue_security_software | 8.69785912e-05 | 2 |
| Scareware | 8.69785912e-05 | 2 |
| Spyware | 8.69785912e-05 | 2 |
| Bliss_(virus) | 8.69785912e-05 | 3 |
| Zlob_trojan | 8.69785912e-05 | 3 |
| Computer_poker_players | 8.69785912e-05 | 3 |
| SCA_(computer_virus) | 8.69785912e-05 | 3 |
| Melissa_(computer_virus) | 8.69785912e-05 | 3 |
| Mimail | 8.69785912e-05 | 3 |
| Blaster_(computer_worm) | 8.69785912e-05 | 3 |
| Back_Orifice | 8.69785912e-05 | 3 |
| Virus_hoax | 8.69785912e-05 | 3 |

In ASURA, we choose multiple choices question as the question type to conduct the adaptive training. In particular, multiple choices question providing relatively competitive incorrect answers, which not only help the user to learn effectively but also keep the learning process from boring [31]. To be more specific, the stem of the multiple choices question is in a straight form of *What is concept*?, where the *concept* is the keyword topic from the learner that they choose to learn. And multiple choices question generated for each concept used in this paper is stored in JSON format as follows:

```
question =
{
'id': concept,
'body': 'What is ' + concept ? ,
'choices': choices [incorrect1, incorrect2, incorrect3, correct],
'ans': ans
}
```

In multiple choices question, the correct answer is coming from the concept itself, and the incorrect answers are derived from those concepts on the same level of the keyword concept from the Computer security hierarchical concept map. Moreover, the idea of generating incorrect choices and correct choices from definition materials is same, and both choices are in a straight form of "*...is concept definition.*", replacing the keyword concept in concept definition with "...". And we using Python Regular Expression to do the replacement in exact text processing. Part of the core query statement of definition in this paper is as follows:

```
SELECT DISTINCT ?definition
WHERE{
⟨http://dbpedia.org/resource/concept⟩ dbo:abstract ?definition.
FILTER (LANG(?definition) = 'en')
}
```

## 6 Adaptive cybersecurity awareness training

In this section, we summary steps of adaptive cybersecurity awareness training and the updating algorithm applied in this paper.

ASURA provides adaptive training that targeted to each individual learner by allowing learner to choose training topic that they want to understand. Each individual training consists of a number of small quizzes that made up by four multiple choices questions. The number of quizzes is determined by the learner's knowledge, the learning speed and the size of subconcept map. The size of subconcept map varied according to the chosen keyword from learner. For instance, concept Spyware is the leaf node of the Computer security concept map. Thus the size of the Spyware subconcept map is only one. In another case, Malware subconcept map including 20 concepts can make training time-consuming. Under this circumstance, we set the threshold for the number of training concepts to six, which can be modified easily in ASURA. In consequence, we selected six concepts from subconcept map according to the PageRank value that calculated in Sect 4.1. Those concepts with higher PageRank value have priority for training. For instance, Tab. 4 showed the selected six concepts for Malware training.

Moreover, similar to conventional adaptive training systems, ASURA contains a straightforward version of learner model to keep track of the learner's understanding. The learner model is kept updated by an updating algorithm. The updating algorithm also brings concept map and learner model together to determine what training content should be given next. The core statement of updating algorithm is as follows:

1. Initiation: Selecting six concepts by traversing the subconcept map level by level, and concept with higher PageRank value priori to be selected.
2. Learner model update: Update the learner model after each quiz according to the feedback understanding from the learner.
3. Generate next quiz: Selecting training concepts by combining the learner model with the subconcept map. Repeat training incorrectly answered concepts based on the feedback from the learner. If the number of unhandled concepts is larger than four, concept with higher PageRank priori to be selected.

ASURA conducts repeat training until the learner handled all selected concepts. For instance, when we conduct a Malware training for a learner who has little background of cybersecurity, the learner would finish the Malware training in four quizzes. Tab. 5 $\sim$ Tab.8 represented four learner models in this practical training. Moreover, Fig. 5 showed part of training quiz that this exemplified learner would take.

Furthermore, ASURA can fully interact with the learner in a command line interface. On the purpose of a good user interface and convenient management, we also combine our system with CyLMS which provided remarkable advantages in training content representation and closer integration with the LMS. More specifically, we first convert JSON format questions to YAML format, and then using CyLMS convert yaml file to SCORM format [21]. Finally, training contents will be displayed in Moodle, a free and open-source learning management system for visualization and interaction.

**Table 4** Selected training concepts related to the Malware

| **Concept** | **PageRank** | **Level** |
|---|---|---|
| Malware | 0.00309623 | 0 |
| Computer_virus | 0.00050839 | 2 |
| Winwebsec | 0.00029768 | 2 |
| Computer_worm | 0.00023484 | 2 |
| AIDS_(Trojan_horse) | 0.00023484 | 2 |
| Linux_malware | 0.00016091 | 2 |

```
********** Quiz 1 **********
******************************************** 1
What is malware?
(a): ...  is the process of creating a new personal identity or alias for an existing person.

(b): ... is a category of malware that targets the users of Windows operating systems and produces fake claims as genuine anti-
malware software, then demand payment to provide fixes to fictitious problems.

(c): A ... (also known as a watch house, guard building, guard booth, guard shack, security booth, security building, or sentry
 building) is a building used to house personnel and security equipment.

(d): ... short for malicious software, is any software used to disrupt computer operations, gather sensitive information, gain
access to private computer systems, or display unwanted advertising.

********************************************
d
Well done! Your answer is correct
******************************************** 2
What is computer virus?
(a): A ... in its narrow sense is an impression left by the friction ridges of a human finger.

(b): A ... ... is a type of malicious software program ("malware") that, when executed, replicates by reproducing itself (copyi
ng its own source code) or infecting other ... programs by modifying them.

(c): ... is the analysis of the physical characteristics and patterns of handwriting purporting to be able to identify the writ
er, indicating psychological state at the time of writing, or evaluating personality characteristics.

(d): ... (from Greek ἄνθρωπος anthropos, "human", and μέτρον metron, "measure") refers to the measurement of the human individu
al.

********************************************
b
Well done! Your answer is correct
******************************************** 3
What is winwebsec?
(a): ... is a category of malware that targets the users of Windows operating systems and produces fake claims as genuine anti-
malware software, then demand payment to provide fixes to fictitious problems.

(b): ... is the analysis of the physical characteristics and patterns of handwriting purporting to be able to identify the writ
er, indicating psychological state at the time of writing, or evaluating personality characteristics.

(c): A ... in its narrow sense is an impression left by the friction ridges of a human finger.

(d): ... (from Greek ἄνθρωπος anthropos, "human", and μέτρον metron, "measure") refers to the measurement of the human individu
al.

********************************************
c
Sorry, that is incorrect!
This is the correct answer:  ... is a category of malware that targets the users of Windows operating systems and produces fake
 claims as genuine anti-malware software, then demand payment to provide fixes to fictitious problems.
******************************************** 4
What is computer worm?
(a): A ... in its narrow sense is an impression left by the friction ridges of a human finger.

(b): A ...  is a standalone malware computer program that replicates itself in order to spread to other computers.

(c): ... is the analysis of the physical characteristics and patterns of handwriting purporting to be able to identify the writ
er, indicating psychological state at the time of writing, or evaluating personality characteristics.

(d): ... (from Greek ἄνθρωπος anthropos, "human", and μέτρον metron, "measure") refers to the measurement of the human individu
al.

********************************************
c
Sorry, that is incorrect!
This is the correct answer:  A ...  is a standalone malware computer program that replicates itself in order to spread to other
 computers.
Malware
Computer_virus
Winwebsec
Computer_worm
+----------------+--------------+
|    Concept     | Check answer |
+----------------+--------------+
|    Malware     |      ✔       |
| Computer_virus |      ✔       |
|   Winwebsec    |      ✘       |
| Computer_worm  |      ✘       |
+----------------+--------------+
********** Quiz 2 **********
```

**Fig. 5** Part of training example about the Malware

**Table 5** Learner model for quiz 1

| Concept | Learner understanding |
|---|---|
| Malware | ✓ |
| Computer_virus | ✓ |
| Winwebsec | × |
| Computer_worm | × |

**Table 6** Learner model for quiz 2

| Concept | Learner understanding |
|---|---|
| Winwebsec | ✓ |
| Computer_worm | ✓ |
| AIDS_(Trojan_horse) | × |
| Linux_malware | × |

**Table 7** Learner model for quiz 3

| Concept | Learner understanding |
|---|---|
| AIDS_(Trojan_horse) | ✓ |
| Linux_malware | × |

**Table 8** Learner model for quiz 4

| Concept | Learner understanding |
|---|---|
| Linux_malware | ✓ |

## 7 Evaluation

In this section, we show evaluation results of ASURA. At first, we evaluate the hierarchical concept map building time. Next, we present evaluation of the hierarchical concept map coverage, continued with validation of functionality provided by ASURA. Finally, evaluation on training effectiveness is provided.

### 7.1 Concept map building time evaluation

In Sect. 3, we described the construction process for the hierarchical Computer security concept map from DBpdeia with timely updating. However, the construction time may increase drastically as the size of concept map increases. Moreover, construction time is an essential factor for system performance that can significantly impact to user experience. As a demonstration, Tab. 9 shows the total time (execution time and the Internet accessing time) for concept map construction with various depth. The time for execution remained on a slow growth while the time for accessing the Internet growing rapidly. The average execution time for those seven concept maps is 1.432918 seconds, while the average time for accessing the Internet is 329.938169 seconds. Especially, when on the relatively large depth, the time for accessing the Internet is almost 779 times slower than the execution time as we observed from Tab. 9.

In ASURA, considering the time issue and effectiveness, we extracted subconcept from the Computer security concept map for the next practical adaptive training. For evaluation, we randomly choose seven different keywords from seven levels of the Computer security map. The result in Tab.11 showed that as depth growing, the total time for the subconcept map building remained stable, as the average time of subconcept map building is 0.51459295 seconds, thus it shows construction time 643.9480 times deduction using our method. Combine the reservation from Tab. 9, we can conclude that the main factor of the hierarchical concept map building time is the Internet accessing time and the subconcept map is a wise method for ASURA.

**Table 9** Concept map building time

| Depth | Concepts number | Execution time(s) | Internet accessing time(s) |
|---|---|---|---|
| 1 | 23 | 0.01110610 | 0.54707691 |
| 2 | 126 | 0.07438310 | 16.10524294 |
| 3 | 331 | 0.35431300 | 80.08356489 |
| 4 | 618 | 0.97561000 | 196.80745818 |
| 5 | 884 | 1.83946000 | 368.93737889 |
| 6 | 1347 | 2.63489700 | 531.09750199 |
| 7 | 2640 | 4.14064700 | 1115.98895925 |

**Table 10** Subconcept map building time

| Concept | Building time(s) | Depth |
|---|---|---|
| Authentication | 0.42738413 | 1 |
| Computer_viruses | 0.45832491 | 2 |
| Malware_in_fiction | 0.47378087 | 3 |
| Malware | 0.56676579 | 4 |
| Authentication_methods | 0.47324681 | 5 |
| Authentication | 0.42738414 | 6 |
| Computer_security | 0.77526402 | 7 |

### 7.2 Concept map coverage

In this section, we assess the coverage of Computer security concept map compared with the ESET Cybersecurity Awareness Training project and the CompTIA Security+ Study Guide.

ESET Cybersecurity Awareness Training advertised that they teach almost everything that the employees need to understand to help make our company's cybersecurity safe [32]. And ESET provides training from five aspects: threats overview, password safety, Internet protection, email protection, and preventive measures. This paper summarized keywords from ESET Cybersecurity Awareness Training project and compared with the concepts from concept map in this research. Tab. 11 shows the matching results with ESET Cybersecurity Awareness Training project, and we found matchings related to each aspect in Computer security map built in Sect. 3. Moreover, one advantage of our method is that we have more detailed and ample concepts for practical training, compared with ESET Cybersecurity Awareness Training project. For instance, the topic Malware in ESET only contains Malware itself, while we can get 20 related concepts.

The CompTIA Security+ is a necessary security certification for IT professionals, and the CompTIA Security+ study guide is a book specially prepared for security technologies who want to earn the Security+ certification[33]. CompTIA Security+ study guide has 12 Chapters, which provides both knowledge base and skills range from physical security and software security. We also summarized the main topics and keywords from that book and then compared with our Computer security concept map. Tab. 12 shows the matching results with this professional IT security book. As a result, we identified all matches in our concept map except two concepts, Securing the Cloud and Security Administration. In fact, there is a concept named Cloud computing security existed in DBpedia Articles dataset. However, our concept map did not include it, as we constructed the hierarchy concept map from the DBpedia Categories dataset, and there is no concept named Cloud computing security existed in Categories. The linking concepts of Cloud computing security in Categories are Computer security and Cloud computing. Firstly, for the concept Computer security, in Sect. 4.2, when we linking the Computer security from Categories to Articles, even with multiple linked concepts, we only chose the most ap-

proximate concept: Computer security in Articles. Secondly, for another linked concept Cloud computing, which is not the descendant concept of Computer security, thus it is not in the built concept map. Considering the fact that CompTIA Security+ is an overall and professional book for IT specialist, our matching results are quite competitive.

### 7.3 Adaptive training functionality evaluation

Furthermore, we evaluate ASURA from the functionality aspect by comparing it with the traditional adaptive training system. Tradition adaptive training systems usually contain four models: Expert model, Learner model, Instructive model, and Instructional environment. Tab. 13 compares the tradition adaptive learning system with the adaptive system prototype and our ASURA system. In generally, ASURA implemented all the primary functions of the four modules. Moreover, it can dynamically and timely update the training materials from DBpedia, as to the expert model. Then for the learner model, traditional ones often contain various personal information, such as age, educational background, etc. in contrast, although the learner model of ASURA only includes one kind of information which is the learner's understanding of the question. Next for the instructive model, the updating algorithm in ASURA combines the processed concepts and feedbacks from the learner to provide the consequent questions. Last for the instructional environment, ASURA provided the learner with a command line interface for the full interaction, with a semi-interacted interface using Moodle.

| ESET concepts | | Concept map concepts | Match |
|---|---|---|---|
| threats overview | malware | malware | ◯ |
| | viruses | compute_viruses ,antivirus_software | ◯ |
| | worms | compute_worms,email_worms | ◯ |
| | trojans | AIDS_(Trojan_horse) | ◯ |
| | ransomware | ransomware | ◯ |
| | rootkit | rootkit | ◯ |
| | spyware | spyware | ◯ |
| | social engineering | engineering_failures,reliability_engineering,software_engineering_costs... | ◯ |
| password safety | password | password_authentication,password_managers,password_cracking_software | ◯ |
| | access | access_control | ◯ |
| | authentication | authentication | ◯ |
| Internet protection | | Internet_security,Internet_privacy_software | ◯ |
| email protection | | email_worms, email_authentication,email_hacking | ◯ |
| preventive measures | spam filter | spam_filtering | ◯ |
| | password manager | password_managers | ◯ |

**Table 11** Concept map coverage vs. ESET Training topics

| **CompTIA Security+** | **Concept map concepts** | **Match** |
|---|---|---|
| Managing risk | IT risk management | ◯ |
| Monitoring and Diagnosing Networks | Computer network security, network analyzers, virtual private networks | ◯ |
| Understanding Devices and Infrastructure | Firewall software,ripple gateways | ◯ |
| Identity and Access Management | Identity management, access control | ◯ |
| Wireless Network Threats | Rogue software | ◯ |
| Securing the Cloud | | — |
| Host, Data, and Application Security | Data security | ◯ |
| Cryptography | Cryptograph | ◯ |
| Threats, Attacks, and Vulnerabilities | Cyberattacks, cryptographic attackst,computer viruses, malware, computer worms, rootkit, adware, spyware,DOS viruses,domain hacks | ◯ |
| Social Engineering and Other Foes | Engineering failures,reliability engineeringaccess control | ◯ |
| Security Administration | | — |
| Disaster Recovery and Incident Response | Back up,intrusion detection systemsspam filtert,port scanners,identity management | ◯ |

**Table 12** Concept map coverage vs. CompTIA Security+ guide topics

**Table 13** System feature evaluation: tradition adaptive learning system vs. system prototype

| **Tradition adaptive system** | **Adaptive system prototype** |
|---|---|
| 1. Expert model | ✓✓training materials:<br>concept map in Section 3<br>question set in Section 5<br>timely updated, extended |
| 2. Learner model | ✓information of the learners:<br>learner model in Section 6<br>contains understanding of the questions |
| 3. Instructive model | ✓✓combines previous two models<br>to provide next question |
| 4. Instructional environment | ✓user interface or training platform<br>fully interacted command line interface<br>semi-interacted with Moodle |

## 7.4 Adaptive training effectiveness evaluation

Finally, we conducted a series of Malware adaptive training to check that whether ASURA can fulfill the purpose of improving people's cybersecurity awareness. This training consisted of 6 training concepts, as shown in Tab. 4. The training concepts were related to the keyword Malware, and the training continued until each learner understood all the related concepts. In total, 8 learners were recruited for participating in this training effectiveness survey. There were 4 learners had cybersecurity background, other 4 did not. The number of quizzes in Fig. 6 revealed the number of small quizzes a learner took to finish the training, with a backline in Fig. 6 indicating the minimum times. And Fig. 7 showed the error count for each learner to take the Malware training. In summary, the average quiz number for the learner with cybersecurity background is 2.25, while the average quiz number for the learner without cybersecurity background is 3.5. It is reasonable that there is a gap between learners with cybersecurity backgrounded and learners without cybersecurity backgrounded. It should be noted that even without advanced text processing method, the learner can still fully understand the questions and made the right choices after training.

Furthermore, for assessing the training effectiveness, we retest those 8 learners in Fig.8 31 days later. As Hermann Ebbinghaus hypothesized the famous forgetting curve in 1885 which demonstrated how the memory of data declines over time when there is no attempt to reinforce it [34]. The retest results showed significant cybersecurity knowledge improvement of the learners. All cybersecurity backgrounded learners finished the Malware training in minimum quiz count, while for those non-cybersecurity backgrounded learners, the quiz count declined. Especially, the quiz count of the learner 6 decreased from 5 to 3. Therefore, in conclusion, the proposed ASURA system was shown to be helpful for learners to improve their cybersecurity awareness effectively.

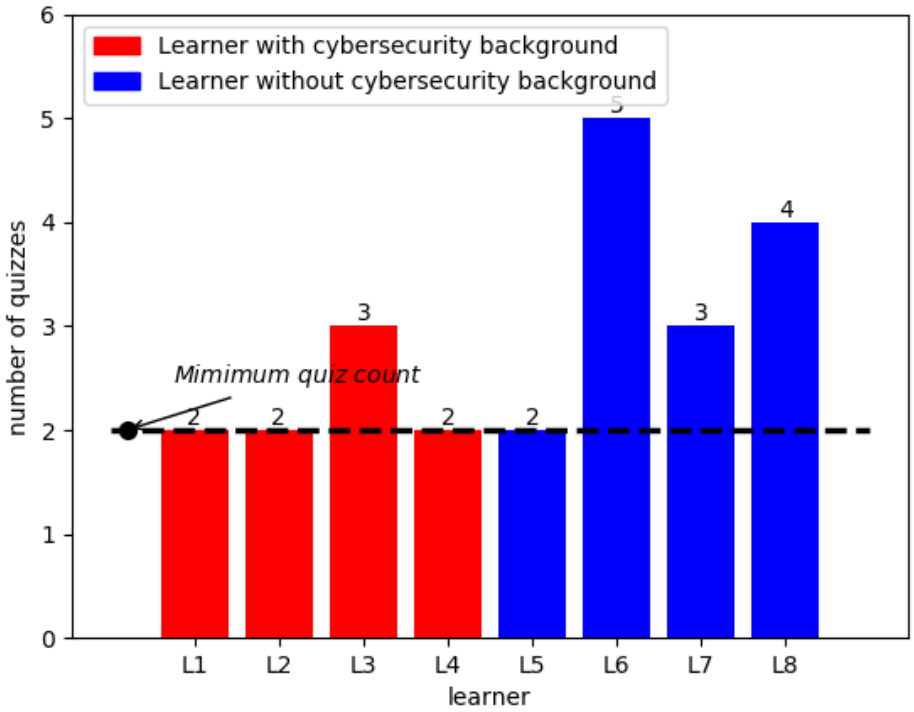


**Fig. 6** The quiz count for each learner to take the Malware training

## 8 Conclusion

In this paper, we implemented an adaptive awareness training system ASURA and solved the three considerations proposed in Sect. 1. Firstly, we used SPARQL querying constructed a Computer security hierarchical concept map from the LOD database DBpedia, dynamically and timely updated. Along with depth growing, the number of concepts increase rapidly and the time for accessing Internet growing fast. In this sense, we proposed the method of extracting subconcept map from the whole Computer security concept map to save time. In conclusion, training contents is up to date with a considerably wide coverage as the timely updated Computer security concept map. Secondly, we employed a series of processing on the built concept map, including relevance estimation and filtering. To be more specific, we define the importance of a concept node as the number of linked concept nodes going out of this concept node. For calculating the importance, we adopt PageRank algorithm on the concept map. By substituting the pages by concept nodes, we can then rank the importance of the concept nodes effectively. Thus, we could decide

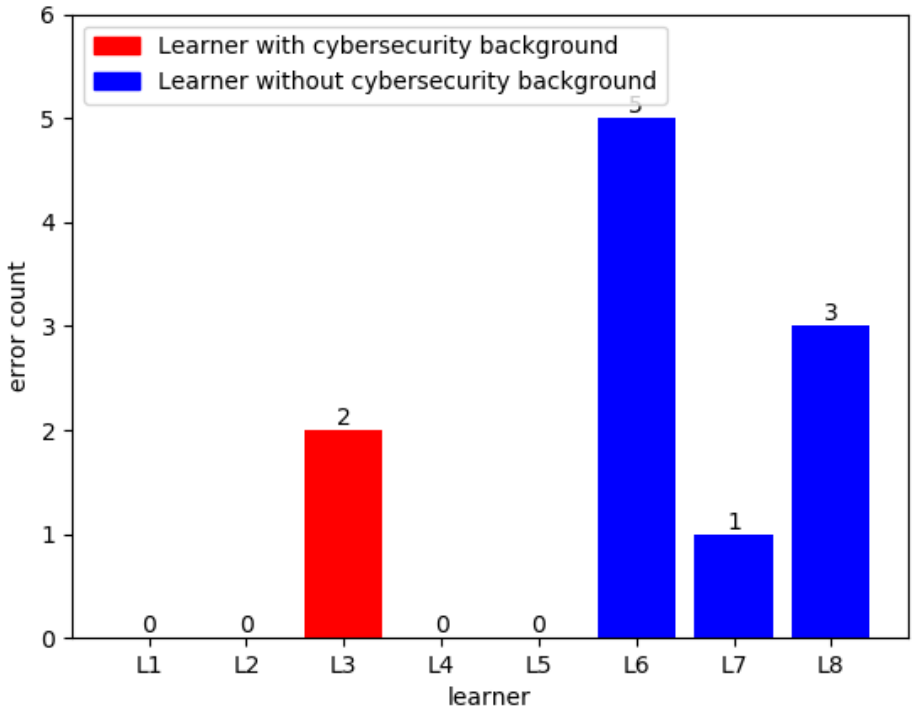


**Fig. 7** The error count for each learner to take the Malware training

the priority of training contents to make the training more efficient. Then we filtered out undefined and unrelated concepts for practical training. Thirdly, we proposed a simple method to deliver the training contents by online adaptive training. ASURA provided fully interaction with learners, and generating training contents with respect to the will of learners, which personalizes each training for individual learners. We use a set of technologies and principles generating multiple choices questions for training. Then we combined learner model and updating algorithm together deciding next training contents for learners.

In the future, this work could be extended in many aspects. As mentioned before, DBpedia is interlinked with other kinds of open datasets. Thus we can get much further information and it is possible to access more cybersecurity-related datasets. In this paper, we using SPARQL to query data from DBpedia, while SPARQL has limitations as a simple query language compared to those programming language. However, some interesting extended SPARQL research had been done on missing features, such as recursion. In the future, extended SPARQL language may be used to optimize the construction strat-

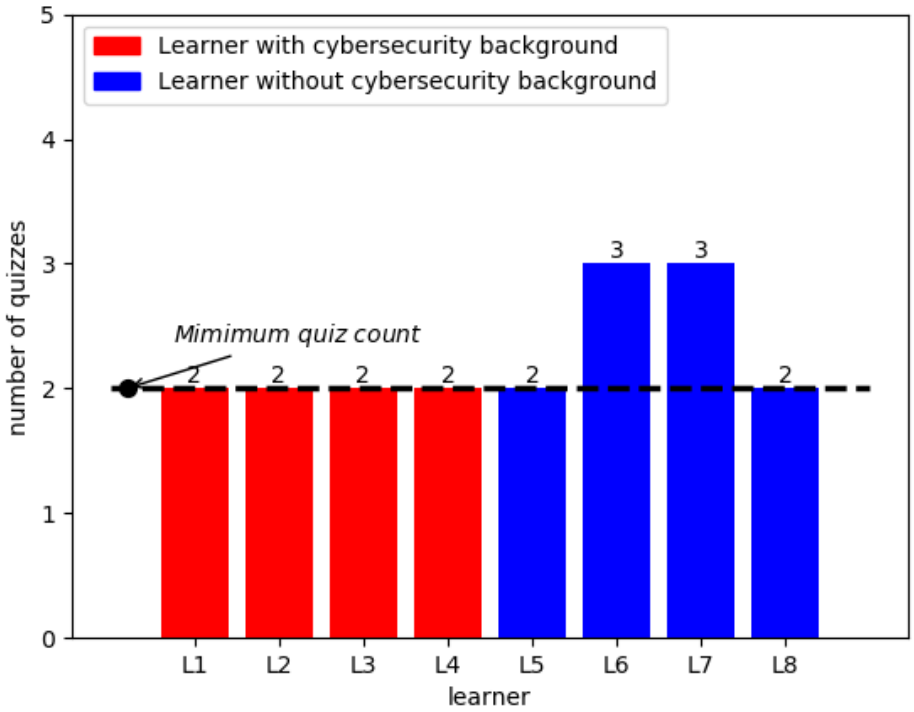


**Fig. 8** Retest: The quiz count for each learner to take the Malware training

egy and reduce the Internet accessing time of the system. We using Python Regular Expressions for text processing in this papaer. Furthermore, in the future, Natural Language Processing techniques can be used in text processing to generate more type-diverse questions for practical training. We applied a straightforward version learner model in this paper that only includes learner's understanding to the knowledge. In the future, the learner model can be extended to more complex one which contains kinds of information of the learners, such as domain knowledge, interests, preferences, goals, tasks, background, etc. The adaptive training strategy can be improved by introducing more factors to decide the next training contents, such as psychometric theory, learning habits, etc. ASURA provided fully interaction with learners in command line interface, future work can be done to improve user interface. For instance, providing fully interaction with learners in Moodle, by introducing other tools, CyLMS for example.

**Acknowledgements** This work was supported by JSPS KAKENHI Grants Number 17K00478 and 17K00479.

## References


1. MediaPRO. 2016 privacy security awareness report. 2016.
2. MediaPRO. 2018 privacy security awareness report. 2018.
3. Jeff Goldman. An urgent need for security awareness training: 30 percent of employees don't know what phishing is. June 2017.
4. FraudWatch. What is cyber security awareness training and why is it so important? Decemeber 2018.
5. Duncan Ki-Aries and Shamal Faily. Persona-centred information security awareness. *Computers & Security*, 70:663 – 674, 2017.
6. Fadi A. Aloul. The need for effective information security awareness. 2012.
7. Jemal Abawajy. User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, 33(3):237–248, 2014.
8. Jemal Abawajy. User preference of cyber security awareness delivery methods. *Behaviour & Information Technology*, 33(3):237–248, 2014.
9. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5:1–22, 2009.
10. Charles R.Kelley. What is adaptive training? *Human Factors*, 11(6):547–556, 1969.
11. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154 – 165, 2009. The Web of Data.
12. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N.Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Soren Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015.
13. DBpedia. About dbpdedia. 2019.
14. Sergey Brin and Lawrence Page. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735. Springer-Verlag, 2007.
15. Eric Prud'hommeaux and Andy Seaborne. Sparql query language for rdf. 01 2007.
16. Pablo Mendes, Max Jakob, and Christian Bizer. DBpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 1813–1817, Istanbul, Turkey, May 2012. European Languages Resources Association (ELRA).
17. Loc Nguyen. Learner model in adaptive learning. 2008.
18. R.S. Shaw, Charlie C. Chen, Albert L. Harris, and Hui-Jou Huang. The impact of information richness on information security awareness training effectiveness. *Computers & Education*, 52(1):92 – 100, 2009.
19. Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi. *Advances in Intelligent Tutoring Systems*. Springer, Berlin, Heidelberg, 2010.
20. Maria Bada, Angela M. Sasse, and Jason R. C. Nurse. Cyber security awareness campaigns: Why do they fail to change behaviour? *ArXiv*, abs/1901.02672, 2014.
21. Razvan Beuran, Dat Tang, Zheyu Tan, Shinobu Hasegawa, Yasuo Tan, and Yoichi Shinoda. Supporting cybersecurity education and training via lms integration: Cylms. *Education and Information Technologies*, 06 2019.
22. E.J. Miller. An introduction to the resource description framework. *Journal of Library Administration*, 34:245–255, 12 2001.
23. W3C. Serializing sparql query results in json. hhttps://www.w3.org/TR/rdf-sparql-json-res/.
24. TopBraid Composer. Property 'has broader'.
25. Navneet Kaur and Diksha Garg. Analysis of the depth first search algorithms. 2012.
26. Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. MIT Press, 1990.
27. S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Seventh International World-Wide Web Conference (WWW 1998)*, 1998.
28. T. H. Haveliwala. Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796, July 2003.

29. TopBraid Composer. Property dct:subject.
30. Strikingloo. Fuzzywuzzy: How to measure string distance on python. 2019.
31. David Nicol. E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, 31(1):53–64, 2007.
32. ESET. Eset authorize training center. https://www.eset.com/us/cybertraining/.
33. Emmett Delaney and Chuck Easttom. *CompTIA Security+ guide*. The name of the publisher, 7 edition, 2018.
34. Jaap M. J. Murre and Joeri Dros. Replication and analysis of ebbinghaus ' forgetting curve murre. 2015.