Adaptive Security Awareness Training Using Linked Open Data Datasets

1610412 TAN ZHEYU

Along with the extraordinary development of technologies, people are getting more and more cyberrelated, and their daily life is exposed to kinds of cyberattacks, such as password attacks, malware attacks and so on. Cybersecurity is no longer an issue discussed only between the professionals or technologists, but it is also closely related to ordinary people. Wombat Security Technologies conducted a survey revealed that ransomware is an unknown concept to nearly two-thirds of employees. Almost 95% of cybersecurity attacks are due to human error. As much as 54% of companies have experienced one or more attacks in the last 12 months, and the number rises every month. A considerable amount of money spent on cybersecurity to protect people, company and organization from cyberattacks. Expensive and sophisticated systems won't play much good without the human factotr which is the main vulnerability in cybersecurity. It has great significance to give people cybersecurity awareness training. The innovative idea of this research is to apply concept map for cybersecurity awareness training. For adaptive cybersecurity awareness training, two issues can't be neglected, as follows.

1. Firstly, what concepts should be used to teach people in awareness training?
2. Secondly, what specific questions should be chosen?
3. Last but not least, how to teach people efficiently?

1 *Concept map construction*

This research proposes a way to construct Computer security concept map from the LOD database DBpedia, dynamically and timely updated, and extracted sub concept map from this big concept map.

DBpedia was constructed by extracting data from Wikipedia which has extensive topic coverage, and it is possible to get much further training materials since it interlinked with other kinds of open datasets. The knowledge in DBpedia is in RDF graph which can be accessed by public SPARQL endpoint. This research uses SPARQL querying data from DBpedia and returns results in JSON format to build the concept map. The query in this research uses an important property: $A\ skos:broader\ B$ which represent hierarchical

relationships between two concepts, and B has a more general and broader meaning than A. Broader concepts are typically rendered as parents in a concept hierarchy. The query strategy is important if we aimed at building an efficient and useful concept map for cybersecurity awareness training. In this research, we reach the descendent concepts of Computer security with Breadth-first algorithm in level order. The descendent concepts of Computer security are enormous. This research shows several Computer security maps with various depth.

In the course of building a concept map, along with depth growing, the number of nodes/concepts increase rapidly and the time for accessing the Internet growing immediately, especially, when on the relatively large depth, the time for accessing the Internet is almost 779 times slower than the execution time. Timely update the big concept map make the cybersecurity materials is fresh. Constructed sub concept map from the big one save much time in the practical training.

2 *Relevance estimation and filtering on concept map*

This research proposed ways to process the built concept map. Employ the PageRank algorithm to calculate the importance of each concept node and do the filtering on concept map for the next adaptive training.

This research constructed the hierarchy concept map from the DBpedia Categories. The sub concept map generated from the Computer security map may still be large, for example, the Malware sub concept map contains 54 concept nodes which may not all useful or so important for the training. Even they are all useful ideally; we still need to decide which concepts have higher priority for Malware adaptive training. This research estimates the concept importance/relevance by applying PageRank algorithm on the Computer security map. Google uses the PageRank algorithm 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. In this research, a concept map is a set of interlinked concept nodes. We assume that the importance of a concept node is determined by the number of linked concept nodes going out of this concept node. The concept node has higher PageRank value considered has higher importance/relevance for adaptive training.

Next, we need to employ filtering on concept map since not every concept has a definition in DBpedia Articles and so related to the training keyword. At first, this research directly checks each concept has a definition or not. But in the late of this research, when we do the exact example with Computer worms, we found out that the Computer worms do not have a definition, but

the Computer worm does have a definition. Without any doubt, no matter Computer worm or Computer worms are useful for awareness training. This unexpected result led us to improve the filter algorithm by linking concept in Categories with the concept in Articles first. The link from Articles to Categories is represented as $dct : subject$. After linking, take concept Malware in Categories as an example, it has multiple results in Articles. Using a NPL library: fuzzywuzzy to find the most approximate string matching of the keyword. If the keyword has a linking concept in Articles, it has a definition in Articles. There is another case needed to be considered, that is, not every concept is relevant to practical training. For example, concept Digimon is in the concept map built from the keyword Malware (Digimon is the grandchildren concept of Malware). But it is an instance of a Game class, it is irrelevant to the cybersecurity awareness training. We filter this kind of concept based on DBpedia class and property.

3 *Using concept map conduct adaptive training*

We proposed a simple way to do adaptive training and implemented an adaptive awareness training system prototype. The processed concept map combined with the simple learner model provided the idea of the adaptive training. Question creation and text processing made training system into actual practice.

Using the filtered concepts to generate questions. This research provided multiple-choice questions. Creat questions using the definition queried from Articles. The stem of the question is in a straight form of $What\ is\ concept$?, where the concept is the given keyword from the learner. The correct answer to the question is coming from the concept itself. The incorrect answers to the question are coming to form the concepts that on the same level of the keyword concept on the Computer security tree. Both choices are in a straight form of "$...is\ concept\ definition$.", replace the concept in concept definition with "..." by using python regular expression operations to do the text processing.

The processed concept map combined with the simple learner model provided the idea of the adaptive training. The learner model present in this research using a straightforward version to interact with the learner and adaptive awareness learning system, which reveals the understandings and misunderstanding to correct knowledge. In pratical training, the system prototype in this research set a threshold of the number of the training concepts to six, this threshold can be modified easily.

Update algorithm bring concept map and learner model together to determine which training content should be given next. In this research, adaptive awareness training contains several small quizzes. Each quiz consists of four questions — the count of quizzes determined by the size of the sub concept map and the knowledge of the learner. In initialize of the first quiz, the update algorithm traverse the sub concept map level by level. In each level, the concept with higher PageRank is the priority to be selected. And after each quiz, update the learner model based on the feedback results from the learner. Then, update algorithm select training concepts based on the learner model and the sub concept map again, the correctly answered concept should not appear in the next quiz, until the learner answered all questions right.

Tradition adaptive learning system usually divided into Expert model, Learner model, Instructive model, and Instructional environment. The adaptive system prototype in this research implemented all the primary function of the four modules. For the expert model, the prototype can dynamically and timely update the training materials and extend the training materials. For the learner model, this prototype employs a straightforward one to track the learner. For instructive model, this prototype combines the processed concept and feedbacks from learner to provide the next question. As the survey did in research background, it is innovative to conduct adaptive training in this way. For the instructional environment, this research prototype provided the learner with a command line interface for full interact.