# Concept Map Building from Linked Open Data for Cybersecurity Awareness Training

Zheyu Tan[1], Shinobu Hasegawa[2] and Razvan Beuran[1]

[1]School of Information Science, Japan Advanced Institute of Science and Technology
[2] Center for Graduate Education Initiative, Japan Advanced Institute of Science and Technology
{jotan, hasegawa, razvan} @ jaist.ac.jp

**Abstract:** With the extraordinary progress made in technology, serious problems have arisen in terms of cybersecurity. There are a large quantity and multiple types of cyberattacks in people's daily life. It has great significance to give people cybersecurity awareness training. But the need to determine on what topics to train people's awareness in cybersecurity comes very first since there are huge cybersecurity materials in the micro-world.

This paper will build a concept map from LOD that can be used for awareness training in cybersecurity. This is done by using SPARQL to query useful data from the LOD database DBpedia to get related data nodes. Then we use the Page Rank Algorithm to calculate the importance of each data node. The more important data nodes have higher priority for cybersecurity awareness training.

**Keywords:** LOD, RDF, SPARQL, concept map, Page Rank Algorithm

## 1. Introduction

Along with the rapid development of technologies, people are getting more and more cyber-related and their daily life are exposed to kinds of cyber-attacks. Cybersecurity is no longer an issue discussed only among the professionals or technologists, but it is closely related to common people. In a survey carried out in 2016 by MediaRRO (specialized in cybersecurity and data privacy employee awareness programs) revealed that 88% US employees lack the awareness needed to prevent common cyber incidents [1]. It is urgent to improve people's cybersecurity awareness and it has great meaning to develop a cybersecurity awareness training.

For cybersecurity awareness training, there are two issues can't be neglected. Firstly, what concepts we should give people to do awareness training and the relevance of concepts used to train people. Secondly, the training contents should be updated timely. The need to determine on what topics and the relevance of topics to train people's awareness in cybersecurity comes very first since there are huge cybersecurity materials in the micro-world. In this paper, we aimed at solving these two problems.

We propose a way to build concept map from the Linked Open Data (LOD) database DBpedia. LOD is sharing machine-readable interlinked data on the Web. DBpedia is one of most famous, frequently updated LOD datasets coming from Wikipedia by extracting data from it. DBpedia has a wide topic coverage, more than 1.95 million resources included, and it is possible to get much further information since it interlinked with other kinds of open datasets. We could dynamic get much related and timely updated cybersecurity materials form it.

This paper employs Page Rank Algorithm to calculate importance of each concept node on concept map to conduct awareness training later. The data nodes have higher importance have higher priority for cybersecurity awareness training.

The remainder of this paper provides a brief background introduction on LOD, RDF (Resource Description Framework) and SPARQL in section 2. In section 3, we explain how to build a concept map from LOD database DBpedia using SPARQL. In section 4, we employ Page Rank Algorithm to calculate the importance of each data node from concept map. In section 5, how to use concept map to give people cybersecurity awareness training will be discussed.

# 2. Preliminaries

This section is a brief introduction to LOD, RDF and SPARQL.

## 2.1 Linked Open Data

World Wide Web was first invented by Tim Burners-Lee in 1989 and it was defined as a system of interlinked hypertext documents that runs over the Internet. Web 1.0 was considered as a “read-only” Web since the user has rarely interaction with the website. The term Web 2.0 was first introduced by Darcy DiNucci in 1999 and became popular by Tim O’Reilly in late 2004. Web 2.0 was defined as a platform where ordinary users can communicate and collaborate by using social software applications, such as Skype, Flickr, YouTube and so on. Web 2.0 was a version of Social Web while Web 3.0 was a version of Semantic Web. The term “Semantic Web” refers to W3C’s vision of the Web of Linked Data. Linked Data can be used for sharing machine-readable interlinked data on the Web, making data understandable to humans but also to machines. Berners-Lee founded the W3C to oversee kinds of standards, and the Semantic Web is also built on these W3C standards: the RDF data model, the SPARQL query language, the RDF Schema and OWL standards for storing vocabularies and ontologies. Berners-Lee introduced a couple of rules which known as the “Linked Data principles” in 2006[3]:

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things.

Linked Open Data is a blend of Linked Data and Open Data which is linked and uses open sources. It breaks down the barriers between different data format and sources. W3C extending the Web by publishing open data as RDF and having RDF links between various data. There are more than 31 billion RDF triples and thousands of datasets on the Web. DBpedia is the one of most famous datasets coming from Wikipedia by extracting data from it. DBpedia has a wide topic coverage and it has described more than 1.95 million resources. In addition, DBpedia is possible of getting much further information since it is interlinked with other kinds of open datasets.

## 2.2 Resource Description Framework

Resource Description Framework (RDF) was originally introduced as data model and published as a W3C recommendation in 1999. RDF provides a general framework for expressing resources. Resource can be anything that has a unique identifier (URI), range from documents, physical objects to abstract concepts. RDF express data in triples: subject, predicate, object. The subject and object represent two related resources and the predicate represents the relationship between them. This triple has a direction from subject to object and the predicate is also called as property. In short, RDF is a directed, labeled graph format model representing information.

## 2.3 SPARQL

SPARQL is “Simple Protocol And RDF Query Language”. It is a standard language to query graph data expressed as RDF triples. SPARQL is capable of querying required and optional graph patterns with specify conjunctions and disjunctions. Usually, SPARQL query is a set of patterns called basic graph pattern with subject, predicate or object may be a variable and the result of a SPARQL query is a solution sequence. In other words, querying data by SPAQL is a process of finding certain graphs that match required graph patterns.

# 3. Concept Map Construction

This section explains how to build concept map from LOD and discuss the query algorithm applied in this paper.

## 3.1 DBpedia Query via SPARQL

This section first use SPARQL querying data from DBpedia and build concept map based on the given keyword. The query will use an important property: skos:broader which the query is based on. This property relates a concept to another concept that is more general in meaning. For example, <A> skos:broader <B> means B has a more general and broader meaning than A. Broader concepts are typically rendered as parents in a concept hierarchy (tree).

SPARQL engines can usually return results in different types, for example, XML format, JSON format and CSV format. In this paper, JSON format as chosen, since the libraries used to process JSON objects are available in most programming languages and it's convenient for the Page Rank Algorithm implementation in this paper. The concept map is in JSON format and to be frankly display it, we will use D3 (D3 or D3.js is a JavaScript library for visualizing data using web standards) to visualize the concept map.

## 3.2 Query Strategy

The query strategy is important if we aimed at building efficient and useful concept map for cybersecurity awareness training. The building strategy is constituted of the following three query methods.

**Method 1**: Given the keyword, query the children concepts and the grandchild concepts of the keyword concept. If error occurs (no children concept or grandchildren concept exits), jump to Method 2.
Figure 1 is the concept map built based on the given keyword: Computer security. If the keyword is Computer worms, error occurred since Computer worms do not have any grandchild concept. Then we jump to Method 2 according to the query strategy.

**Method 2**: Only query the children concepts of the keyword concept. If error occurs (no children concept exits), jump to Method 3. Figure 2 is the concept map built based on the given keyword: Computer worms. If the keyword is Password cracking software, error occurred since concept Password cracking software do not have any child concept. Then we jump to Method 3 according to the query strategy.

**Method 3**: Find sibling concepts of the keyword concept by finding the parent concepts first and then children concepts of parent concepts. This method is optional since it received related concepts.
Figure 3 is the concept map built based on the given keyword: Password cracking software

Built concept maps can be stored locally as training material caches for cybersecurity awareness learning. In another way, concept map can be constructed on learner's demands (learner input keyword) and timely updated. The concept map in Figure 1 has 128 nodes and the time to build it is around 1 minutes (including Internet access).

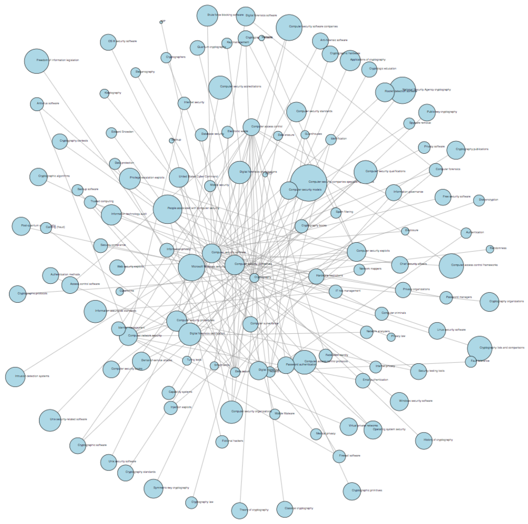

Figure 1. Concept map built from method 1 (Keyword: Computer security).

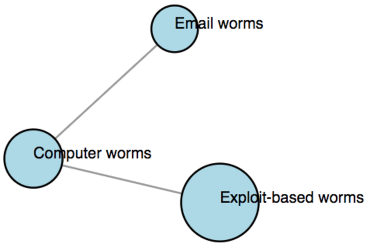

Figure 2. Concept map built from method 2 (Keyword: Computer worms).

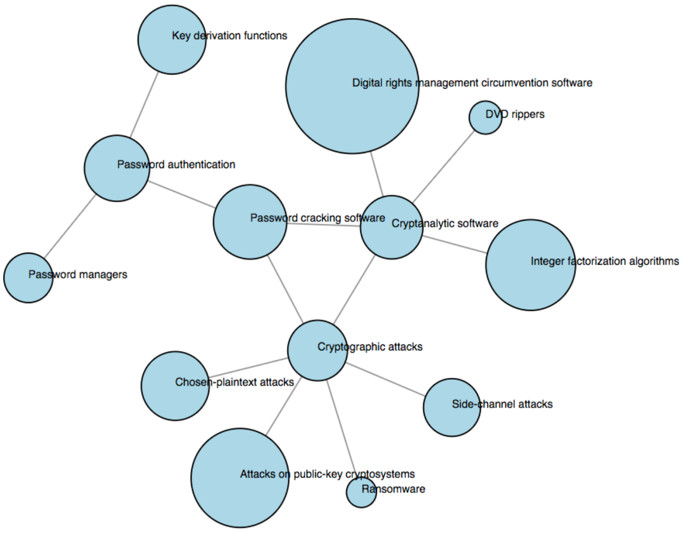

Figure 3. Concept map built from method 3 (Keyword: Password cracking software).

# 4. Concept Relevance Estimation

The continuous of this section introduced Page Rank Algorithm and employed it to the built concept map in section 3.

## 4.1 Page Rank Algorithm

Page Rank algorithm is one of core methods that Google uses to determine the relevance or importance of a page. Page Rank algorithm is defined as follows [7]:

*We assume page A has pages $T_1 \dots T_n$ 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 page A. The PageRank of a page A is given as follows:*

$$PR(A) = (1 - d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad （1）$$

Lawrence Page and Sergey Brin give an Intuitive Justification in their published paper [7]. They consider PageRank as a model of user behavior that user random suffer the Internet. The users visit a page with a certain probability is its PageRank. This probability is given by the number of links on that page and the PageRank of a page is divided by the number of links on that page. The damping factor d is the probability that the random surfer jumping to another random page at each page, so the probability is represented as a constant $(1 - d)$ in the above definition.

There is another version of Page Rank algorithm that Lawrence Page and Sergey Brin published in another paper. In this second version, the probability of random surfer a page is weighed by the total number of web pages. The page's PageRank is the actual probability for a random surfer reaching that page after clicking many links. Then the PageRanks form a probability distribution over the all web pages and the sum of all page's PageRank will be one. In fact, these two versions have no fundamentally difference. The second version of PageRank of a page A is given as follows:

$$PR(A) = \frac{(1 - d)}{N} + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad （2）$$

Where $N$ is the total number of all pages on web. In this paper, the second version (2) will be used since we want to take the total number of concepts into account.

Table 1: Final PageRanks (Keyword: Data security)

| **Concept** | **PageRank value** |
|---|---|
| Data security | 0.3180509599185708 |
| Computer access control | 0.10508589035403851 |
| Backup software | 0.06747620069590507 |
| Electronic waste | 0.058073778281371696 |
| Fault tolerance | 0.03926893345230496 |
| Backup | 0.029866511037771604 |
| Data erasure | 0.02046408862323824 |
| Information governance | 0.02046408862323824 |
| Database security | 0.02046408862323824 |
| Authentication | 0.011061666208704874 |
| Fault-tolerant computer systems | 0.011061666208704874 |
| Free backup software | 0.011061666208704874 |
| Uninterruptible power supply | 0.011061666208704874 |
| Access control software | 0.011061666208704874 |
| Computer access control frameworks | 0.011061666208704874 |
| Federated identity | 0.011061666208704874 |
| Electronic waste in Europe | 0.011061666208704874 |
| File deletion | 0.011061666208704874 |
| Computer access control protocols | 0.011061666208704874 |
| Computer security | 0.011061666208704874 |
| Digital rights management | 0.011061666208704874 |
| Error detection and correction | 0.011061666208704874 |
| Identity management | 0.011061666208704874 |
| Password authentication | 0.011061666208704874 |
| Backup software for DOS | 0.011061666208704874 |
| Backup software for Mac OS | 0.011061666208704874 |
| System image | 0.011061666208704874 |
| Data breaches | 0.011061666208704874 |
| Smart cards | 0.011061666208704874 |
| Backup software for Linux | 0.011061666208704874 |
| Backup software for Windows | 0.011061666208704874 |
| Computer recycling | 0.011061666208704874 |
| Electronic waste in Africa | 0.011061666208704874 |
| Electronic waste in Asia | 0.011061666208704874 |
| Turing tests | 0.011061666208704874 |
| Online backup services | 0.011061666208704874 |
| Disk cloning | 0.011061666208704874 |
| Electronic waste by country | 0.011061666208704874 |

## 4.2 PageRank Calculation

Google use Page Rank 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. In Page Rank algorithm, the ranking of a page is recursively given by the ranking of those pages which linked to it. In this paper, in the same way, the ranking of a concept is recursively given by the ranking of those concepts which related/linked to it. But how do we know the final PageRank value of those concepts since the PageRank of a concept is always given recursively by the PageRank of related concepts. The answer can be found in Lawrence Page and Sergey Brin's paper [7]. We repeat calculations many times until the values stop changing much ($(P_{n+1} - P_n) < \varepsilon$). In this paper, we guess PagePanks of concepts at first and it doesn't matter where you start. When the PagePank calculations converged or reached a fixed point, the normalized probability distribution will be 1.0. According to Lawrence Page and Sergey Brin, around 100 iterations are necessary to get a good approximation of the PageRank values of the whole web. Table 1 is the final calculated importance/PageRank value on Data security map after 48 iterations and it is organized in the descending order. The damping factor is set to 0.85, the maximum number of iteration is set to 100, the epsilon is set to $10^{-6}$.

## 5. Awareness Training

In section 3, given the cybersecurity keyword, we built a concept map from DBpedia. In section 4, we employed Page Rank Algorithm and calculated the importance of each concept node on concept map. In this section, we use previous results to conduct cybersecurity awareness training for people.

*Filtering concepts based on DBpedia class and property.* In previous concept map, 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. This paper discards irrelevant concepts.

*Filtering concepts based on concepts utility.* There is another case needed to be considered, that is, not every concept has definition in DBpedia. On hypothesis is which concept without definition is not important for practical training. This paper discards no definition concept.

*Generating automatically questions.* In section 4.2, we finally generated a list of final PageRanks. We sorted this ranking list in descending order and choose several top concepts to generate questions for training. The concept has higher importance/ PageRank value has higher priority to train people.

So far, after filtering irrelevant and no definition concepts and picking up top important concepts. We generated prompts and correct choices for trainee. Incorrect choices are selected randomly from training concepts and apply simple text processing to make choices. Figure 4 is the overview for this cybersecurity awareness training.

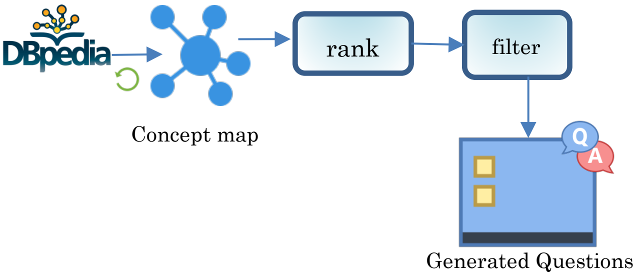

Figure 4. Training process overview.

## 6. Conclusion

This paper contributed to determining training concepts for cybersecurity awareness learning by dynamically constructing a concept map from the LOD database DBpedia, and which can be timely updated. We employed the Page Rank Algorithm on the built concept map and the result, sorted ranking list, can be used to decide the training concepts relevance in cybersecurity awareness training. Finally, this paper is capable of determining the training concepts and their relevance for cybersecurity awareness learning; moreover, training materials can be timely updated.

Next step, we will extend this research from the following aspects: (a) Extend the LOD resources to cybersecurity related databases to get more substantial data for training; (b) Improve query strategy for keywords with small number of children/siblings.

## Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 17K00479.

## References

[1] RH Strategic, Kevin Eike: Report: 88% of Employees Lack the Awareness Needed to Prevent Common Cyber Incidents, October 26, 2016.

[2] Nupur Choudhury: World Wide Web and Its Journey fromWeb 1.0 to Web 4.0, 2014.

[3] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far, 2009.

[4] The linked Open Data from lod-cloud.net, https://lod-cloud.net/, 2018.

[5] RDF 1.1 Primer, W3C Working Group Note, https://www.w3.org/TR/rdf11-primer/#section-Introduction, June 24, 2014.

[6] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd: The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, January 29, 1998.

[7] Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, Volume 30 Issue 1-7, April 1, 1998.

[8] W3C: SPARQL Query Language for RDF, W3C Recommendation15,January,2008,https://www.w3.org/TR/rdf-sparql-query/#ask.

[9] Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary G. Ives: DBpedia: A Nucleus for a Web of Open Data, ISWC/ASWC, 2007).

[10] Max Schmachtenberg, Christian Bizer, Heiko Paulheim: State of the LOD Cloud 2014, Version 0.4, August 30, 2014.

[11] TDWG Terms Wiki: Definition of “skos:broader”, https: //terms.t dwg.org/wiki/skos:broader.