JAIST Repository
https://dspace.jaist.ac.jp/

Title	Semantic-awareness Recommendation with Linked Open Data in Web-based Investigative Learning
Author(s)	丁, 庚
Citation	
Issue Date	2021-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/17109
Rights	
Description	Supervisor: 長谷川 忍, 先端科学技術研究科, 修士(情報科学)

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Master's Thesis

Semantic-Awareness Recommendation with Linked Open Data in Web-Based Investigative Learning

TING Kang

Supervisor HASEGAWA Shinobu

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

February, 2021

Abstract

With the rapid development of information and communication technology in the 21st century, human life has also undergone tremendous improvements. In the field of education, we have considerable expectations and emphasis on the development of the web resources such as Linked Open Data(LOD) which is combined a blend of Linked Data and Open Data. LOD breaks down barriers between different data formats and sources. Web-based investigative learning is one of the learning approaches benefited from.

As the model of Web-based investigative learning provides a platform for learners to create their own learning scenarios by organizing knowledge over the web in self-directed way. This kind of knowledge management activity helps learners to achieve a high cognitive load on investigation. However, it is difficult for learners to discover related concepts among a vast number of unstructured web resources concurrently with a better knowledge construction process. Therefore, this research aims to propose a method to recommend semantic-related concepts with Linked Open Data for learners during the investigation of the web-based investigative learning process.

We proposed a Semantic-awareness Recommendation System which extracting the relevant concepts at different levels from DBpedia. DBpedia is a linked open data project which extracts structured content from Wikipedia. Those structured content represented as Resource description framework(RDF) graph allowed the user to query the relationships and properties of Wikipedia resources semantically. In this work, generating a regulated concept map based on the initial question for the recommendation, three significant elements would be considered:

- Semantic relations: According to the SKOS document, the properties broader and narrower are used to assert a direct hierarchical link between two concepts.
- Node importance: The PageRank algorithm would calculate the importance of the concepts extracted by regulated SPARQL query strategy.
- Content containment: It is based on concept utility. For example, not every concept has definition in DBpedia. The hypothesis is which concept without definition is not important for the recommendation.

For the concepts extraction, we proposed a Regulated Concept Map Generation process by using regulated SPARQL query strategy. We firstly extract Simple Knowledge Organization System(SKOS) Concepts(RDF graph) from DBpedia using SPARQL query. Then, related concepts with semantic relations(Broader-Narrower) would be returned. The essential property: **SKOS:broader** would be used. This property represents a hierarchical relation between concepts. It is important for us to regulate the SPARQL query strategy if we aim to recommend the related concepts at different levels without preventing learners from their self-directed investigation. The regulated concepts map is a collection of entities called nodes, which are concepts that we are going to recommend to learners. Concepts are linked by edges with the property ***SKOS:broader***(Broader) and ***is SKOS:broader of***(Narrower).

Since the PageRank algorithm is generally used as an index to decide the importance of nodes in a directed graph such as the RDF graph. Therefore, the PageRank algorithm is suitable for the concept importance estimation of this work, and we named it as Semantic-aware PageRank. We assume that the importance of a concept node is determined by the number of outbound links on that concept. The probability of random surfer a node is weighted by the total number of nodes in the Regulated Concept Map.

Before updating the recommendation list to learners, we have to filter those concepts which are not important. In this work, we would filter concepts based on the concept's utility. Since not every concept has a definition on DBpedia, the hypothesis that the concept has no definition on DBpedia is not significant for the recommendation.

For the evaluation, the concept importance estimation results would be analyzed by Spearman's correlation coefficient. Spearman's correlation coefficient measures the strength and direction of the association between two ranked variables. we had compared the strength and direction of the association between DBpagerank, User expectation, and Semantic-aware PageRank by Spearman's correlation coefficient. Owing to the finding of analyzing results, Semantic-aware PageRank maintained most serious strength of the association between User expectation. Furthermore, a case study was conducted for testing the hypothesis that using our proposed recommendation system could help learners strengthen the knowledge construction process by discovering semantic-related concepts during Web-based investigative learning. The results evaluated by statistical methods suggest that Semantic-awareness recommendation with linked open data promotes the efficiency of the knowledge construction process.

Acknowledgement

Throughout the writing of this thesis, I have received a great deal of support and assistance.

First of all, I wish to express my greatest appreciation towards my supervisor, Associate Professor HASEGAWA Shinobu, for the patient guidance, encouragement and advice he has provided throughout my time as his student. He treated me well with respect since the first day we meet. I have been extremely lucky to have a supervisor who cared so much about my work, and who responded to my questions and queries so promptly.

I would like to express my gratitude to Associate Professor IKEDA Kokolo and NGUYEN Minh Le. This work has greatly benefited from their valuable comments during the mid-turn presentation.

I would also like to thank all members of HASEGAWA Laboratory for the suggestions they made in the meetings. Each of whom has provided patient advice and guidance throughout the research process. I am so glad to have KOBAYASHI, NAKAGAWA and HORIGUCHI. We made a lot of precious memories throughout my time in JAIST.

Last but not least, I am extremely grateful to my family for their love, caring and sacrifices for educating and preparing me for my future. I am very much thankful to my uncle DING Qiang for the financial help. Without his support, I could never have had the chance to study in Japan.

Contents

1 Introduction **1**
- 1.1 Research Background . . . 1
- 1.2 Problem Statement . . . 2
- 1.3 Thesis Structure . . . 2

2 Research Background **4**
- 2.1 Connectivism . . . 4
- 2.2 Web-based Investigative Learning . . . 4
  - 2.2.1 interactive Learning Scenario Builder (iLSB) . . . 6
- 2.3 Linked Open Data . . . 6
- 2.4 DBpedia . . . 7
- 2.5 Resource Description Framework . . . 9
- 2.6 SPARQL Query . . . 10
- 2.7 Simple Knowledge Organization System . . . 11
- 2.8 PageRank Algorithm . . . 12
  - 2.8.1 The Original Google's PageRank Algorithm . . . 12
  - 2.8.2 The Second Version of Google's PageRank Algorithm . 13
- 2.9 Related Works . . . 13
  - 2.9.1 Adaptive Recommendation for Question Decomposition in Web-based Investigative Learning . . . 13
  - 2.9.2 Relevance between Q-keywords Corresponding to Transition of Interest in Web-based Investigative Learning . 14

3 Semantic-awareness Recommendation System **15**
- 3.1 System Design . . . 15
- 3.2 Regulated Concept Map Generation . . . 16
  - 3.2.1 SPARQL Query and Simple Knowledge Organization System(SKOS) . . . 17
  - 3.2.2 Directional SPARQL Query Strategy . . . 18
- 3.3 Concept Importance Estimation(Semantic aware PageRank) . 21
- 3.4 Filtering . . . 22

4 Evaluation **25**
4.1 Comparison between Semantic-aware PageRank, DBpagerank and User expectation 25
4.1.1 Analysis of Spearman's Correlation Coefficient 28
4.1.2 Analysing Results 29
4.2 Case Study 31
4.2.1 Analyzing Results of First Task 32
4.2.2 Analyzing Results of Second Task 34
4.2.3 Analyzing Results of Post-test Questionnaire 36
4.2.4 Discussion 41

5 Conclusion **43**
5.1 Future Work 43

List of Figures

2.1 The model of Web-based investigative learning[15]. 5
2.2 User interface of iLSB[14]. 6
2.3 The Linked Open Data Cloud[2]. 8
2.4 The overview of DBpedia components[16] 9
2.5 An instance of RDF on DBpedia 10
2.6 An example of retrieving top-10 Universities on DBpedia. . . . 11

3.1 Overview of Semantic-awareness recommendation system[5]. . 16
3.2 Extracting SKOS Concepts(RDF graph) from DBpedia using SPARQL query. 18
3.3 The query for extracting broader concepts of initial Q-keyword. 19
3.4 The query for extracting narrower concepts of initial Q-keyword. 19
3.5 Overview of regulated concept map generation 20
3.6 An instance that how we arrange the range of SPARQL query. 21
3.7 The range and nodes will be applied for the practical case. . . 23
3.8 The SPARQL query for filtering. 24

4.1 Two directional query strategies for the case study. 31
4.2 An example of interlinking between two concepts. 42

List of Tables

2.1 Properties of Simple Knowledge Organization System[5] 12

3.1 Properties of SKOS were employed in this work[5] 17
3.2 Top 10 concepts related to Natural Language Processing sort by Semantic-aware PageRank. 23

4.1 **More general** concepts of **Machine learning** recommended by proposed method and its ranking. 26
4.2 **More specific** concepts of **Machine learning** recommended by proposed method and its ranking. 27
4.3 **More general** concepts of **Smoking** recommended by proposed method and its ranking. 27
4.4 **More specific** concepts of **Smoking** recommended by proposed method and its ranking. 28
4.5 The analysing results of of Spearman's Correlation Coefficient. 30
4.6 Descriptive statistics for two groups in this case study. 34
4.7 Concepts selected by the participants corresponding to the levels. 35
4.8 Descriptive statistics of the participants' perception of the satisfaction. 38
4.9 Descriptive statistics of the participants' perception of the effectiveness. 40
4.10 More general concepts of Machine learning recommended by proposed method and its ranking. 41

Chapter 1

Introduction

1.1 Research Background

With the rapid development of information and communication technology in the 21st century, human life has also undergone tremendous improvements. In the field of education, we have considerable expectations and emphasis on the development of the web resources such as Linked Open Data(LOD)[8] which is combined a blend of Linked Data and Open Data. LOD breaks down barriers between different data formats and sources. Web-based investigative learning[12] is one of the learning approaches benefited from. It allows learners to investigate any topics to learn in a self-directed way.

The web-based investigative learning model included three processes[15]: search for web resources, navigational learning, and question decomposition. Learners need to select suitable and reliable resources against an initial keyword for knowledge construction from a vast number of web resources by themselves. This learning means searching the meaning of the initial keyword and exhaustively investigating many concepts related to the initial question and construct broader and deeper knowledge. By repeating these processes cyclically, learners are expected to create a learning scenario that means turning those unstructured web resources into structured resources to make their knowledge construction process strengthen. However, it is difficult for learners to discover related concepts among a vast number of unstructured web resources concurrently with a better knowledge construction process.

Therefore, this thesis aims to propose a method to recommend semantic-related concepts with LOD for learners during the investigation of the web-based investigative learning process. LOD is a set of structured data interlinking with related ones on the Web. In this work, we use DBpedia[16].

1.2 Problem Statement

In the previous works of web-based investigative learning, Hagiwara et al.[12] proposed an approach to diagnosing insufficiency of a learning scenario created by a learner to extract and recommend Question keyword(Q-keyword) representing sub-keyword to be decomposed using Linked Open Data. The adaptive recommendation makes an effort to help learners who have difficulty constructing a learning scenario, sufficiency decomposing the learning scenario, and observing the relation between Q-keyword during web-based investigative learning. However, when we focus on learners' self-initiative, the recommendations should follow their learning process from the decomposition and the comprehensive concepts that learners are newly interested.

For supporting the long-term learning scenario creation, Yamauchi[23] proposed a method that computes the relevance between two Q-keywords for learners to obtain the awareness between a pair of concepts. His work does provide a good chance for learners to recognize the relation between learning scenarios partially by exploiting LOD. However, the capabilities of LOD was underutilized. It is not representative enough to express the relation between concepts comprehensively.

Several research projects[20][10] focus on semantic path-based ranking using LOD such as DBpedia to generate a ranked recommendation list and tuning the weights of features gathered from DBpedia to increase recommendation accuracy. However, according to the criteria of web-based investigative learning, every learner will create specific learning scenarios in a self-directed way against different Q-keywords.

In order to tackle those issues, we proposed a method to generate the regulated concept map against selected keyword for recommending the relevant concepts at different levels without preventing learners from their self-directed investigation with LOD and Semantic relations between concepts. We also proposed defining concepts related to the initial keyword by employing a Simple knowledge organization system, PageRank algorithm, and filtering strategy against DBpedia.

1.3 Thesis Structure

The structure of this thesis is as follows:

- Chapter 1 : Introduction.
  In this chapter, we introduce the aims of this research, and the problems we face were also stated.

- Chapter 2 : Research Background.
  In this chapter, essential background knowledge for this research would be introduced.

- Chapter 3 : Semantic-awareness Recommendation System.
  In this chapter, we introduce how the recommendation system was designed which included Regulated Concept Map generation, Concept Importance Estimation and Filtering.

- Chapter 4 : Evaluation.
  In this chapter, we aim to evaluate the concept importance estimation and recommendation approach we proposed by statistical methods.

- Chapter 5 : Conclusion.
  In this chapter, we concluded achievements of this work and how could it be improved in the future.

Chapter 2

Research Background

This chapter will introduce significant background knowledge for this research.

2.1 Connectivism

According to the definition of connectivism[19], learning is no longer just a process of personal acquisition of materialized knowledge, but a process of establishing connections to build an individual's internal cognitive network and external social network. Web-based investigative learning[15] provides a platform for learners to investigate the knowledge that resides in the Web needs to be connected. This kind of knowledge management activity helps learners to address those issues of organizational knowledge and transference. It could be seen from previous work[15] that web-based investigative learning improves the efficiency of the knowledge construction process for learners.

2.2 Web-based Investigative Learning

Since the proposed recommending approach is for Web-based investigative learning, the basis of such learning and the cognitive tool named interactive Learning Scenario Builder(iLSB)[14] conducted for promoting question decomposition will be described.

In the previous work of Web-based investigative learning[15], this learning model included three stages(Figure 2.1): searching for web resources/pages, navigational learning, and question decomposition. Firstly, learners would search for web resources with a keyword representing an initial keyword also named Q-keyword. This stage aims to find out appropriate web resources for question investigation. Then, learners could navigate those resources selected

in the previous stage for the knowledge construction process during the navigational learning stage. Meanwhile, they would also extract keywords from navigated resources. Finally, learners could build their own learning scenario by reviewing the knowledge constructed in the navigational learning stage to decompose the Q-keyword into sub-question to be further investigated.

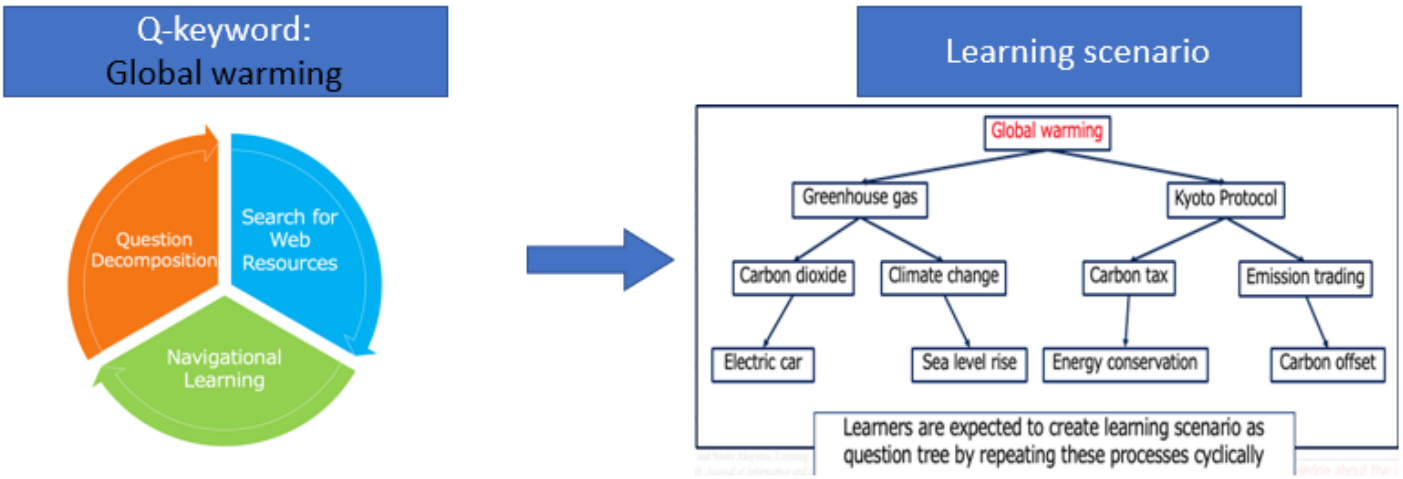

Figure 2.1: The model of Web-based investigative learning[15].

The main feature of web-based investigative learning is to turn those unstructured web resources into structured resources by learners in their self-directed way. By comparing web resources with traditional text resources, text resources are well structured and provide learning scenarios that imply the questions to be investigated and their sequence, such as the table of contents. On the other hand, Web resources are unstructured and do not provide learning scenarios in advance. Therefore, learners need to decompose questions into related ones as sub-questions while constructing their knowledge. It implies that learners are expected to investigate questions in a self-directed way. Meanwhile, learners should create their learning scenarios and construct their knowledge concurrently. As a result, learners would have a high cognitive load on the investigation.

However, it is difficult for learners to discover concepts related to existing learning scenarios and estimate the relevance between a bunch of related concepts. If learners create a new learning scenario with weak relevance between previous scenarios they created, it is difficult for them to strengthen the knowledge construction process. Therefore, the necessity of recommendation should be regarded.

2.2.1 interactive Learning Scenario Builder (iLSB)

In order to scaffold learners' investigative learning process as modeled, a cognitive tool named interactive Learning Scenario Builder(iLSB)(Figure 2.2)[14] has been developed as an add-on for Firefox. iLSB provides scaffolding functions, Searching engine for gathering learning resources, Keyword repository for constructing their knowledge, and Question tree viewer for creating their learning scenario. Owing to previous work finding[15], it has ascertained that iLSB could promote question decomposition in Web-based investigative learning.

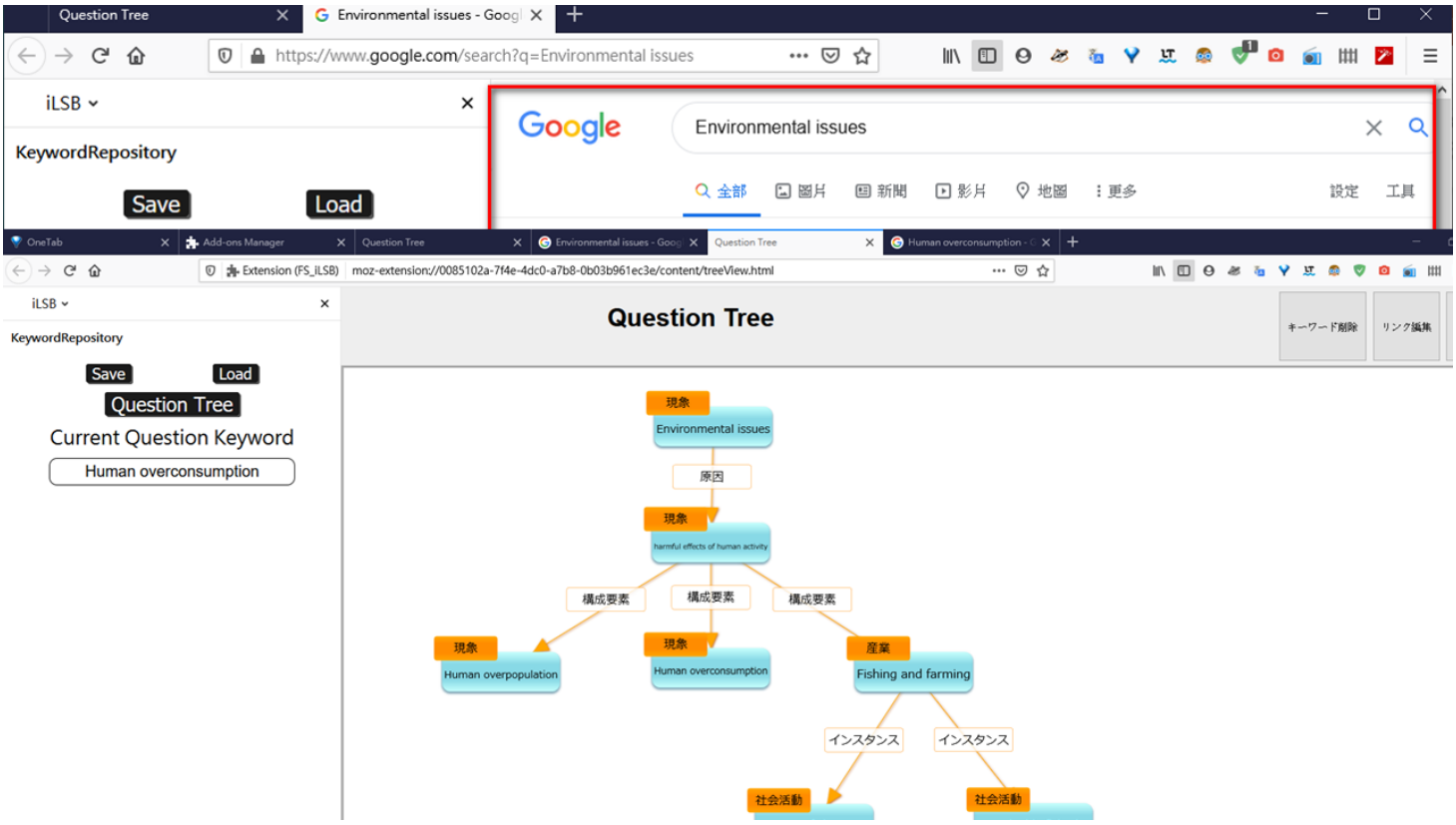

Figure 2.2: User interface of iLSB[14].

2.3 Linked Open Data

In 1989, Tim Berners-Lee invented the first proposal[8] for the World Wide Web. The proposal outlined the principal concepts, and it defined important terms behind the Web, such as describing the Internet as a system of an

interlinked hypertext document. He founded the World Wide Web Consortium(W3C) to maintain the development of those open standards to ensure the long-term growth of the Web. In addition to the classic "Web of documents", W3C also built a technology stack to support a "Web of data" named linked data. The ultimate goal of linked data is to enable computers to do more useful work and develop systems that can support trusted interactions. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS. In 2006, Berners-Lee released the principles of linked data[3]:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information.
- Include links to other URIs. so that they can discover more things.

Therefore, Linked Open Data is a set of structured data interlinking with related ones on the Web that is linked and uses open resources. One remarkable example of a LOD set is DBpedia which extracts structured information from Wikipedia and makes it available on the Web. In this work, we will retrieve relevant concepts from DBpedia for the recommendation.

2.4 DBpedia

Since its establishment in 2007, the DBpedia project[16] has been sustainably releasing large and open data sets, which are extracted from Wikimedia projects(such as Wikipedia and Wikidata[21]). The data has been extracted using a sophisticated software called DBpedia Information Extraction Framework (DIEF) and represented by using the Resource Description Framework(RDF)[4]. In the last few years, the system has received many extensions and fixes from the community, which leads to the creation of a stable release version. Furthermore, by the effort of the W3C Semantic Web Education and Outreach (SWEO) interest group, DBpedia interlinked to a lot of massive Linked Open Data sets. Figure 2.3 shows the Linked Open Data Cloud[2] that DBpedia plays a significant part in Linked Open Data.

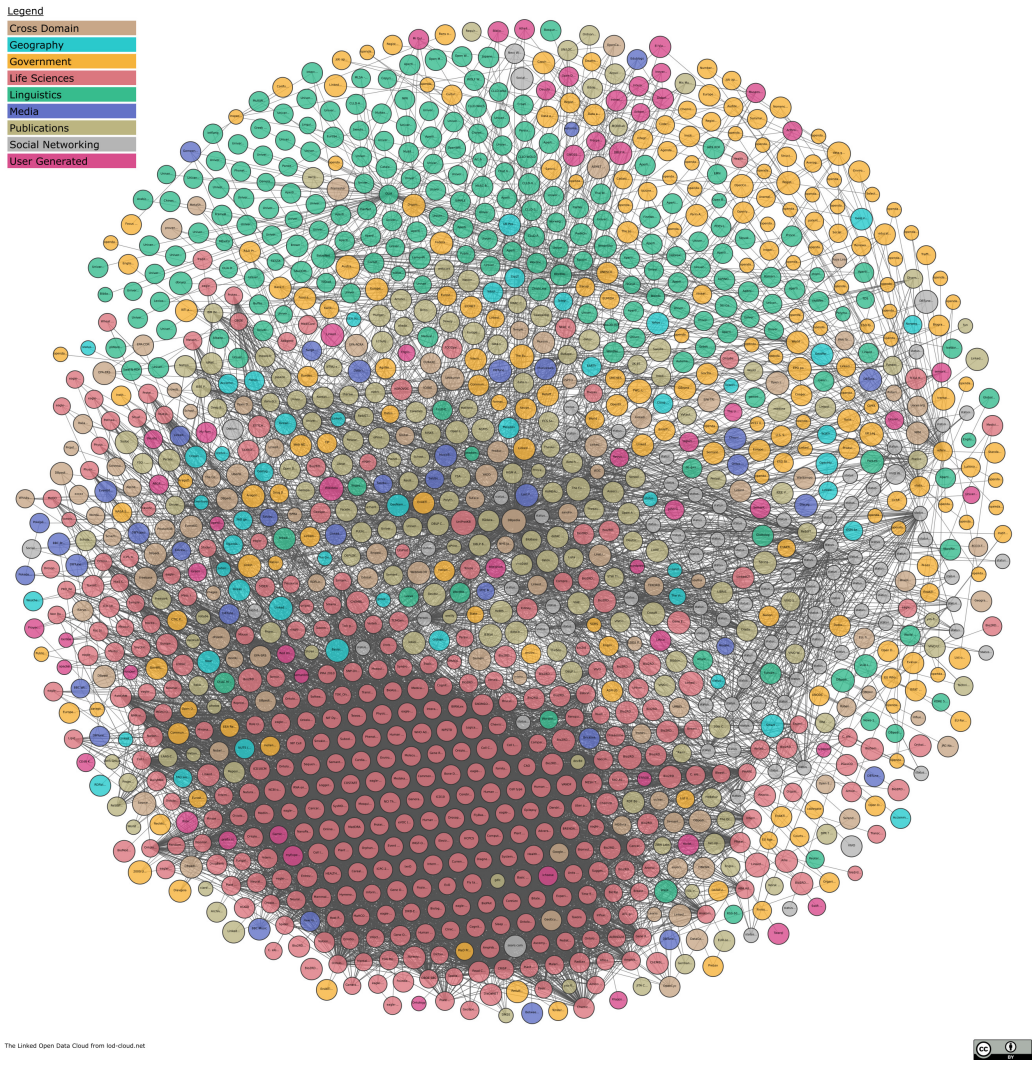

Figure 2.3: The Linked Open Data Cloud[2].

The English version of DBpedia contains 6.0 million entities, of which 4.6 million have abstracts[1]. It means DBpedia has a huge range of subject coverage. Moreover, DBpedia also consists of 5.0 billion pieces of information(RDF triples)[1] extracted from the English edition of Wikipedia. Meanwhile, an information extraction framework that included extraction, clustering, uncertainty management, and query handling was developed by the DBpedia community[16]. Figure 2.4 shows the overview of DBpedia components. It means that it is convenient for us to query those structured data represented by the Resource Description Framework, especially the relationships and properties of information. All in all, for retrieving concepts to

provide a recommendation, DBpedia is a reliable data-set for this work.

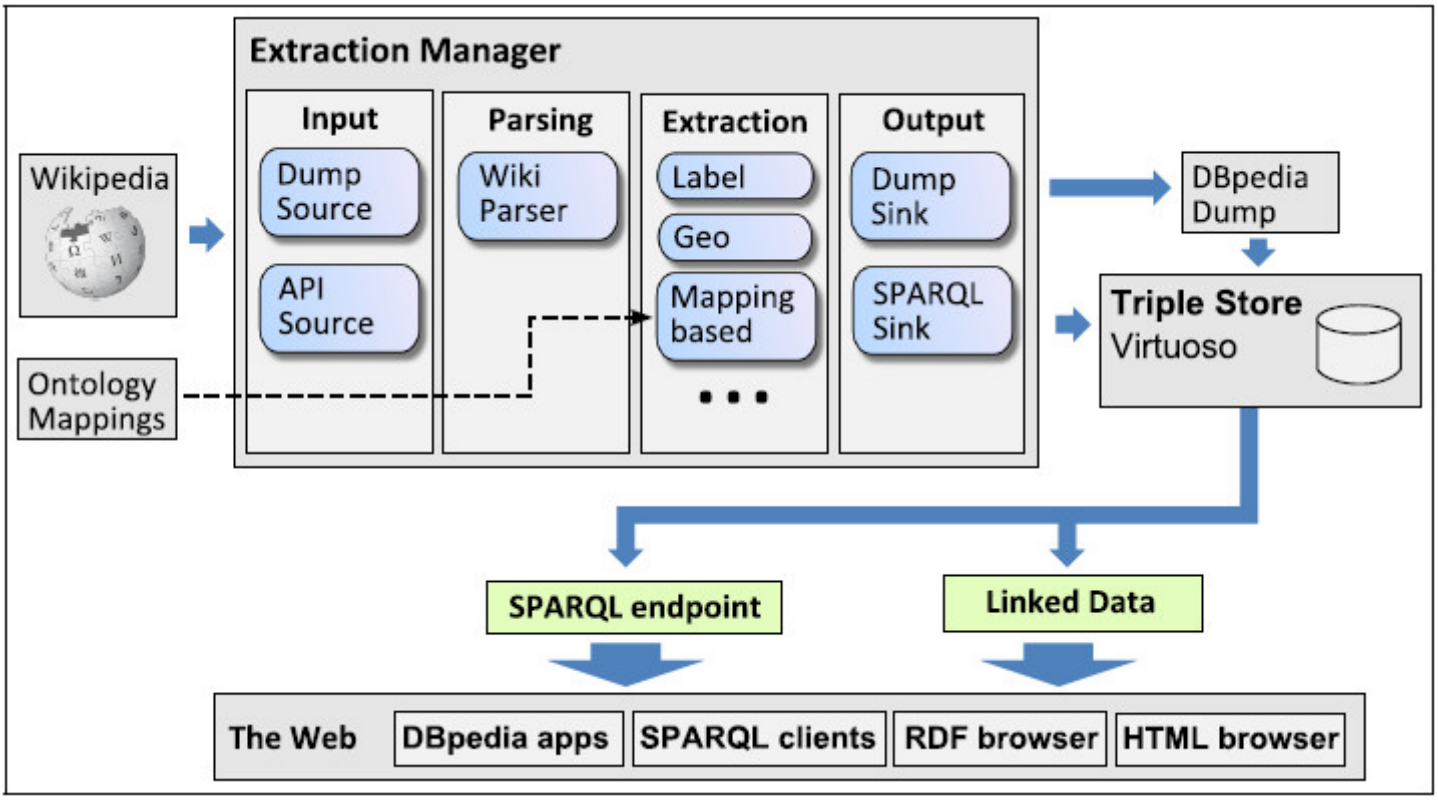

Figure 2.4: The overview of DBpedia components[16]

2.5 Resource Description Framework

Resource description framework(RDF)[4] was conducted by RDF Core Working Group under W3C. RDF data represent a data model for the information over the web as well as DBpedia. The model of RDF data express information in a triple, which included three elements such as subject, predicate, and object. Figure 2.5 shows an instance on DBpedia that the relationship between subject and object is described by predicate which is directional and represented by property.

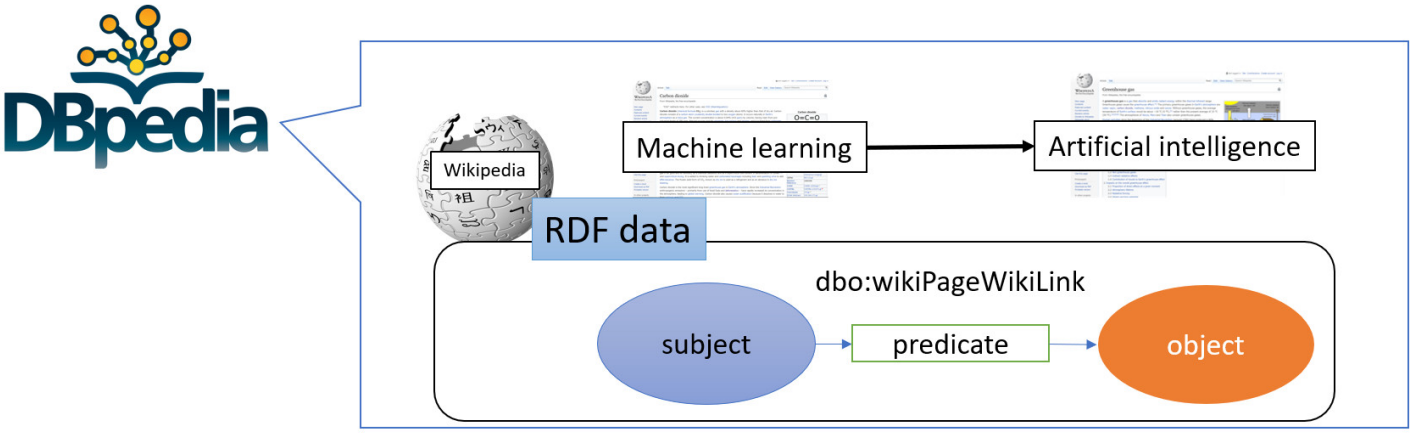

Figure 2.5: An instance of RDF on DBpedia

We could immediately realize that a collection of triples can be represented as a graph data model named as RDF graph[4] which is labeled and directed. This kind of structured data benefited the construction of content that we are going to recommend in our proposed method.

2.6 SPARQL Query

For querying the RDF data, SPARQL Protocol and RDF Query Language (SPARQL) was developed by the W3C RDF Data Access Working Group (DAWG)[6]. It is a standard query language and protocol for RDF graph data. SPARQL query is used for querying required and optional RDF graph patterns with specifying conjunctions and dis-junctions. Generally, in the SPARQL query, whether subject, predicate, or object could be the target variable of the RDF graph data. That is to say, sending a SPARQL query is a process to search the RDF graph data, which matches with required graph patterns. According to DAWG[6], the SPARQL query form included **SELECT**, **CONSTRUCT**, **DESCRIBE** and **ASK**. By the combination with modifiers such as **LIMIT**, **ORDER BY**, **FILTER**, and so on, we could easily query all of the required graph patterns as we need against LOD. By sending the SPARQL query to Public DBpedia SPARQL endpoint[1], we could extract all of the RDF graphs in DBpedia. Figure 2.6 shows an example of retrieving top-10 Universities which are ordered by DBpageank[13].

[1]Public DBpedia SPARQL endpoint: https://dbpedia.org/sparql

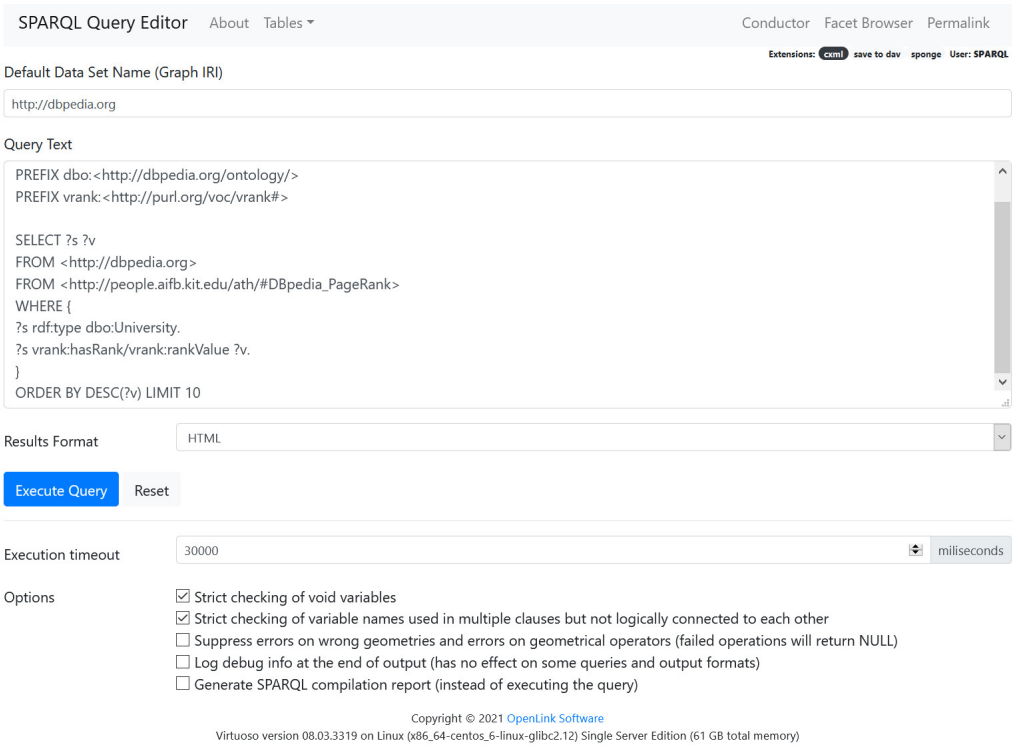

Figure 2.6: An example of retrieving top-10 Universities on DBpedia.

2.7 Simple Knowledge Organization System

Simple knowledge Organization System(SKOS)[5] is a W3C recommendation document that defined a standard data model for sharing and linking knowledge organization systems via the semantic web. The principal element categories of SKOS are concepts, labels, notations, documentation, semantic relations, mapping properties, and collections. It is useful for us to describe the relationship between concepts, such as in a semantic-awareness way. The associated elements are listed in the table below.

Table 2.1: Properties of Simple Knowledge Organization System[5]

SKOS Vocabulary Organized by Theme

Concepts	Labels & Notation	Documentation	Semantic Relations	Mapping Properties	Collections
Concept	prefLabel	note	broader	broadMatch	Collection
ConceptScheme	altLabel	changeNote	narrower	narrowMatch	orderedCollection
inScheme	hiddenLabel	definition	related	relatedMatch	member
hasTopConcept	notation	editorialNote	broaderTransitive	closeMatch	memberList
topConceptOf		example	narrowerTransitive	exactMatch	
		historyNote	semanticRelation	mappingRelation	
		scopeNote			

2.8 PageRank Algorithm

PageRank algorithm is a major algorithm that Google uses to evaluate the relevance or importance of a web page. There are two different versions of the PageRank algorithm were published by Lawrence Page and Sergey Brin in several publications[9][18]. Section 2.8.1 shows the Original Google's PageRank algorithm, and Section 2.8.2 describes the second version.

2.8.1 The Original Google's PageRank Algorithm

Quoting the description of the Original PageRank algorithm published by Page and Brin[9] is given by:

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

Where:

- $PR(A)$ is the Original Google PageRank of page A.
- $PR(T_n)$ is the Original Google PageRank of page T_n which linked to page A.
- $C(T_n)$ is number of the outbound links on page T_n.
- d is a damping factor that could be set between 0 and 1.

- n is the total number of pages that linked to page A

In the Original Google's PageRank algorithm, the importance of a page T is constantly weighted by the number of its outbound links $C(T)$. It means that the more outbound links a page T has, the less page A would be benefited by a link from page T. The PageRank value of page A would be the sum of the inbound links which multiplied by a damping factor d is generally set to 0.85[11].

2.8.2 The Second Version of Google's PageRank Algorithm

For the second version, the PageRank value of page A is as follows[18]:

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

Obviously, these two versions of the PageRank algorithm have no fundamental difference between each other. However, in the second version, it adapts (1 - d)/N instead of (1 - d) where N is the sum of all web pages. It means the probability of a random user surfer a web page is weighted by the total number of web pages. It forms a probability distribution over web pages, and the sum of PageRank value of all pages on the web would be 1.

2.9 Related Works

This work is inspired by the previous researches of web-based investigative learning providing high-quality recommendations and awareness of the relevance between concepts for learners to strengthen their knowledge construction process.

2.9.1 Adaptive Recommendation for Question Decomposition in Web-based Investigative Learning

According to previous works of Web-based investigative learning, Hagiwara et al.[12] pointed out that learners often suffer from question decomposition during Web-based investigative learning. It is difficult for learners to make a sufficient investigation in concurrence with navigation and knowledge construction. Therefore, an adaptive recommending strategy for providing a related sub-question keyword against an initial Q-keyword by extracting the data from DBpedia was proposed. Owing to the finding of this work, it

is ascertained that providing recommendations could promote the question decomposition and elaborate their knowledge constructed. However, the recommending strategy should not only focus on question decomposition but also learners' self-initiative. Therefore, we proposed an approach that recommends the relevant concepts at different levels without preventing them from the self-directed investigation.

2.9.2 Relevance between Q-keywords Corresponding to Transition of Interest in Web-based Investigative Learning

For concerning the transition of interest in web-based investigative learning, Yamauchi[23] pointed out that we should focus on the initial Q-keyword and those concepts learners are newly interested in. He defined three parameters to calculate the relevance between two questions by LOD as follows:

- Question distance: The number of nodes that appear in the shortest path to connect two questions on DBpedia.
- Question similarity: Simpson's coefficient between two sets consisted of related words of each question.
- Question coupling: The number of found elements connecting with question keyword in both directions on DBpedia.

By means of DBpedia, the relevance between a pair of Q-keywords could be calculated. His work breaks down the barriers of evaluating the relevance between different learning scenarios in Web-based investigative learning. By exploiting the capabilities of LOD, we could express the relations between concepts comprehensively over the semantic web. Therefore, for our proposed method, those recommended concepts related to the initial Q-keyword will be defined by three elements which included **Semantic relations**, **Node importance** and **Content containment**. It would be further explained in the following chapter.

Chapter 3

Semantic-awareness Recommendation System

3.1 System Design

The design of the Semantic-awareness recommendation system is illustrated in Figure 3.1. Firstly, the system requests learners to input an initial question for extracting relevant concepts from DBpedia. The initial question could be the concepts existed in the learning scenarios used to create or other keywords that learners are newly interested in. By means of a regulated SPARQL query strategy, we can retrieve related concepts at different levels. The Regulated Concept Map Generation process would be presented in Section 3.2. Secondly, Concept importance estimation will be mentioned in Section 3.3. In this work, we employ PageRank algorithm to estimate the importance of those concepts we retrieved in the previous section. The importance of nodes in generated Regulated Concept Map would be calculated by the PageRank algorithm. Finally, it is significant for us to define the filtering condition before updating the recommendation list to the learner. The filtering strategy is based on the content containment of the concept. Which means that if there is no definition on DBpedia for the concept, we have to filter it. The details of the filtering strategy would be mentioned in Section 3.4. As a result, proposed system updates the recommendation list for learners to continue the navigational learning process.

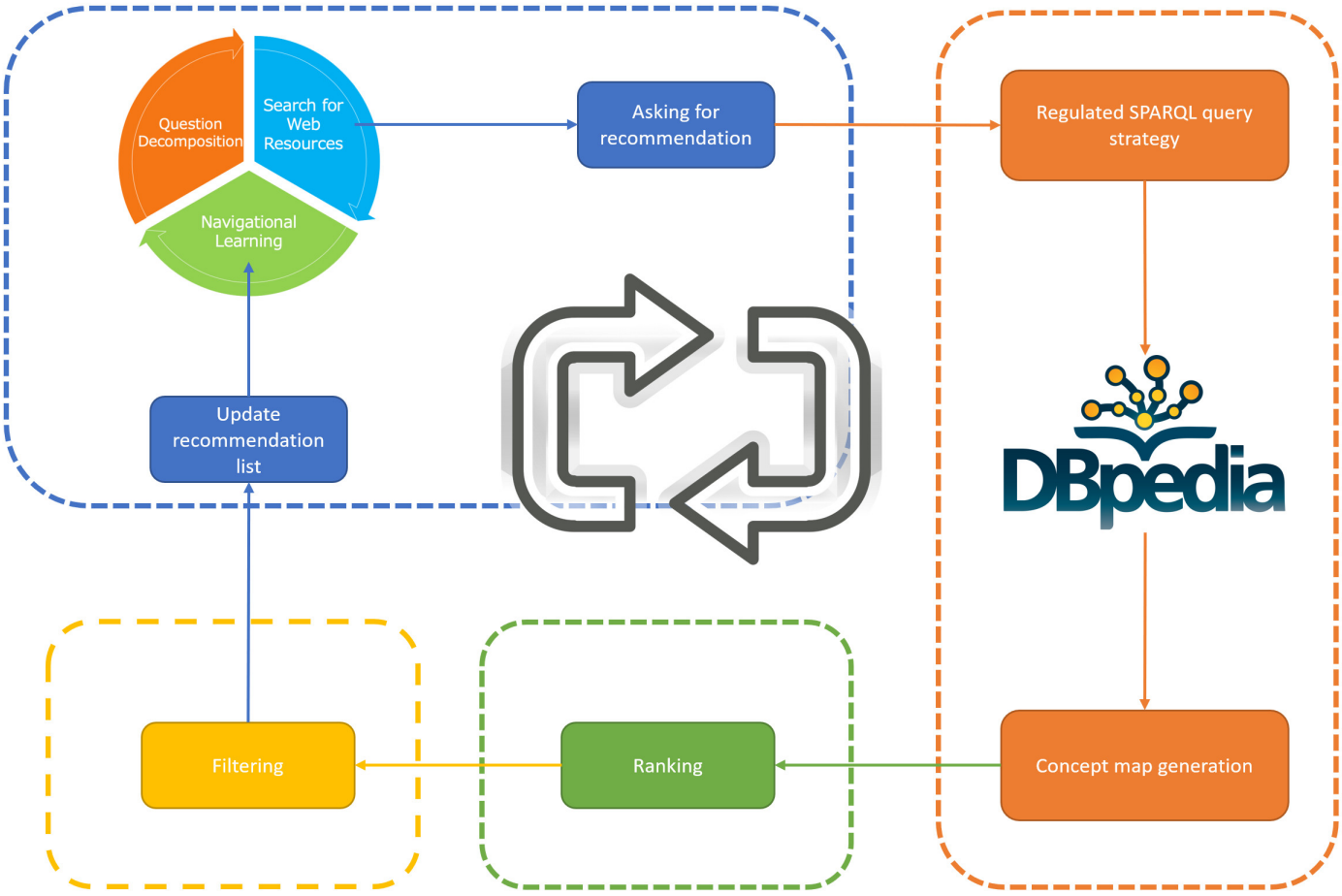

Figure 3.1: Overview of Semantic-awareness recommendation system[5].

3.2 Regulated Concept Map Generation

In this section, we will introduce how the regulated concept map was generated. We extract the relevant concepts at different levels from DBpedia using the SPARQL query. Meanwhile, the regulated SPARQL query strategy was defined. In this work, generating a regulated concept map based on the initial Q-keyword, three significant elements would be considered:

- Semantic relations: According to the SKOS document, the properties broader and narrower are used to assert a direct hierarchical link between two concepts.
- Node importance: The PageRank algorithm would calculate the impor-

tance of the concepts extracted by regulated SPARQL query strategy.

- Content containment: It is based on concept utility. For example, not every concept has definition in DBpedia. The hypothesis is which concept without definition is not important for the recommendation.

3.2.1 SPARQL Query and Simple Knowledge Organization System(SKOS)

DBpedia is a linked open data project which extracts structured content from Wikipedia[16]. Those structured content represented as RDF graph allowed the user to query the relationships and properties of Wikipedia resources semantically. We can either download the entire data set or access it by the public SPARQL endpoint. In this work, accessing DBpedia by public SPARQL endpoint is preferred since DBpedia's dataset will be updated in the future.

As we mentioned in Section 2.7, The important elements **Concepts** and **Semantic Relations** of the SKOS were employed in this work. SKOS concept is defined as RDF resources, and SKOS semantic relations are designed to declare the relationship between concepts within the scheme. The associated elements were employed in this work are listed in the table below.

Table 3.1: Properties of SKOS were employed in this work[5]

SKOS Vocabulary Organized by Theme					
Concepts	**Labels & Notation**	**Documentation**	**Semantic Relations**	**Mapping Properties**	**Collections**
Concept	prefLabel	note	broader	broadMatch	Collection
ConceptScheme	altLabel	changeNote	narrower	narrowMatch	orderedCollection
inScheme	hiddenLabel	definition	related	relatedMatch	member
hasTopConcept	notation	editorialNote	broaderTransitive	closeMatch	memberList
topConceptOf		example	narrowerTransitive	exactMatch	
		historyNote	semanticRelation	mappingRelation	
		scopeNote			

Extracting semantic related concepts from DBpedia, we firstly extract SKOS Concepts(RDF graph) from DBpedia using SPARQL query. Then,

related concepts with semantic relations(Broader-Narrower) would be returned. The essential property: **SKOS:broader** would be used. This property represents a hierarchical relation between concepts. For example, ***A SKOS:broader B*** means B is broader and has more general meaning than A. Narrower follows in the same pattern. Take an initial Q-keyword **Machine learning** as an instance(Figure 3.2). If we want to find out those concepts have broader relation with the initial Q-keyword, we first send the SPARQL query to extract keywords with semantic relations, and the related keywords will be returned.

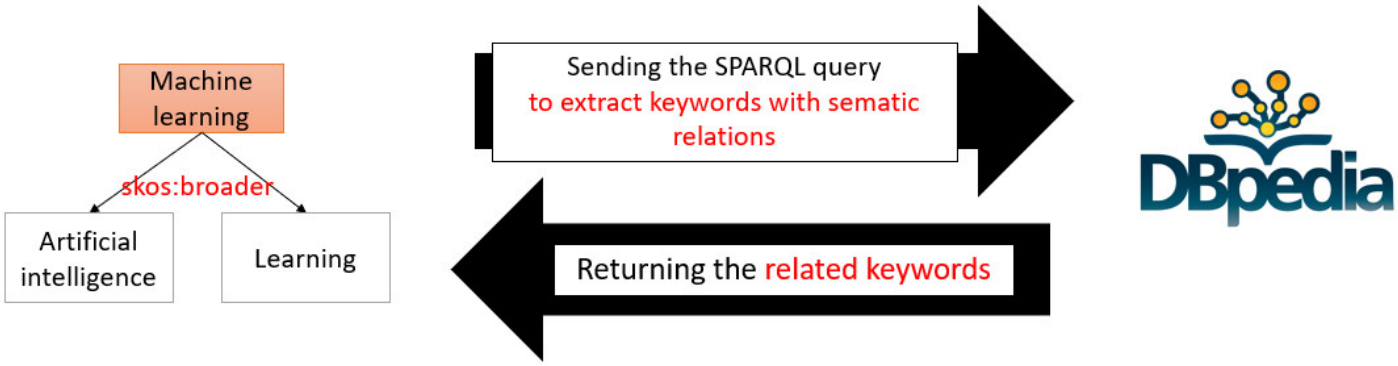

Figure 3.2: Extracting SKOS Concepts(RDF graph) from DBpedia using SPARQL query.

Public SPARQL endpoint could return the query results in different formats such as JSON, CSV, and N-Triples. In this work, we select N-Triples for the result format since it is convenient for us to convert undirected graphs to a directed graph with two directed edges for each undirected edge. N-Triples is a line-based, plain text format for encoding an RDF graph. It is not only for the calculation of PageRank algorithm but also for the ***A is SKOS:broader of B*** relation between concepts narrower of ***A SKOS:broader B***.

3.2.2 Directional SPARQL Query Strategy

It is important for us to regulate the SPARQL query strategy if we aim to recommend related concepts at different levels without preventing learners from their self-directed investigation. The regulated concepts map is a collection of entities called nodes, which are concepts that we are going to recommend to learners. Concepts are linked by edges with the property

SKOS:broader and ***is SKOS:broader of***. The queries asking for broader nodes and narrower nodes used in this research are as follows.

```
sparql.setQuery(f"""
CONSTRUCT {{ {concept} skos:broader ?parent .}}
WHERE {{ {concept} skos:broader ?parent .}}
""")
```

Figure 3.3: The query for extracting broader concepts of initial Q-keyword.

```
sparql.setQuery(f"""
CONSTRUCT {{ ?child skos:broader {concept} .}}
WHERE {{ ?child skos:broader {concept} .}}
""")
```

Figure 3.4: The query for extracting narrower concepts of initial Q-keyword.

By combining the queries above, Figure 3.5 shows all of the semantic related concepts at different levels on DBpedia could be extracted. We need to pay attention to the definition of the PageRank algorithm. If the node has no outbound link, its importance would be 0. Therefore, it is necessary for us to adjust the range of the SPARQL query for the regulated concept map. Figure 3.6 shows an instance that when we focus on the importance of Parent Nodes and Sibling Nodes from Parents Node, the range of SPARQL query should be more in-depth.

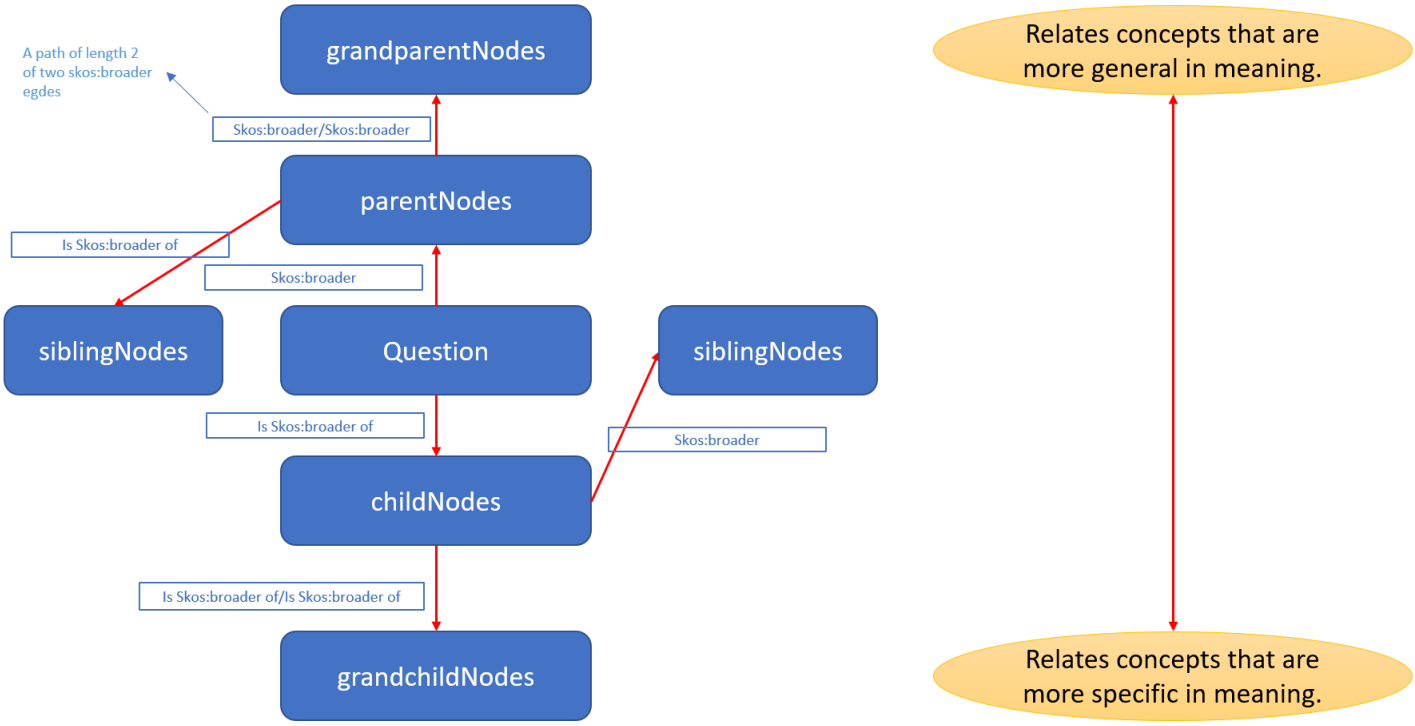

Figure 3.5: Overview of regulated concept map generation

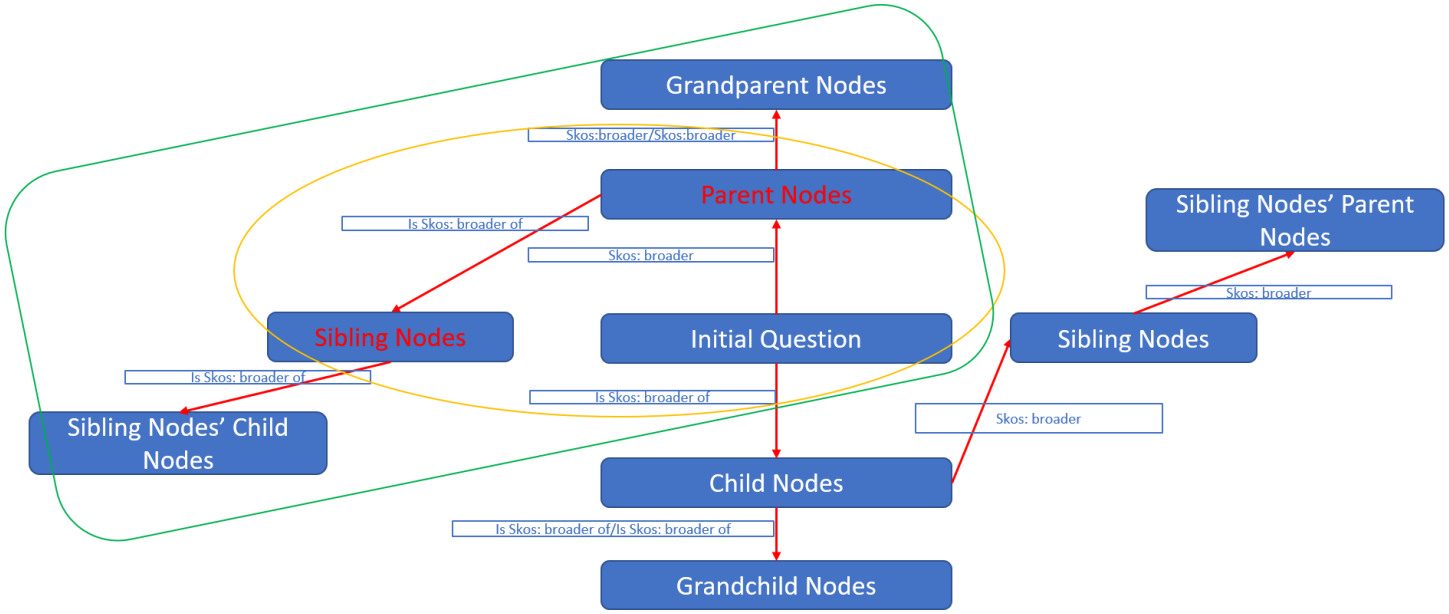

Figure 3.6: An instance that how we arrange the range of SPARQL query.

3.3 Concept Importance Estimation(Semantic aware PageRank)

The original PageRank algorithm was introduced in section 2.8. In the original paper that Lawrence Page and Sergey Brin published, they consider the PageRank algorithm as a model of user behavior who randomly suffer a web page with a certain probability, and the probability is given by the links on that web page. Moreover, due to the previous work finding[13], the PageRank algorithm is also generally used as an index to decide the importance of nodes in a directed graph such as the RDF graph. Therefore, the PageRank algorithm is suitable for the concept importance estimation of this work, and we named it as Semantic-aware PageRank. In this research, a regulated concept map is a set of interlinked nodes, and we defined:

- The number of all nodes in Regulated Concept Map as $|R|$.
- A set of nodes x with the links $\{x, r\} \in E$ where $r \in R$ as B_r.
- The number of links from node x as C_x.

Eventually, we can calculate the PageRank value for all nodes in Regulated Concept Map PR_r based on the equation below:

$$PR_r = \frac{1-d}{|R|} + d \sum_{x \in B_r} \frac{PR_x}{C_x}$$

Where d is a damping factor, which is set as 0.85[11].

We assume that the importance of a concept node is determined by the number of outbound links on that concept. The probability of random surfer a node is weighted by the total number of nodes in the Regulated Concept Map. The equation above forms a probability distribution only over all the nodes in the Regulated Concept Map, and the sum of them would be 1.

In fact, there existed the calculated PageRank value for DBpedia. There are the reasons that we did not generally use the calculated PageRank value on DBpedia for concept importance estimation. The first reason is that it is significant for us to regard learners' knowledge construction process during Web-based investigative learning. Concepts were recommended to guide learners to navigate related concepts with strong relevance between initial Q-keyword. The second reason is that it is essential for us to regard the content containment of those recommended concepts. Concepts such as Time period have certain importance in general cases. However, for the knowledge construction process, the utility of the concept is not important for the recommendation. These reasons concluded the importance of directional SPARQL query strategy.

3.4 Filtering

Before updating the recommendation list to learners, we have to filter those concepts which are not important. In this work, we would filter concepts based on the concept's utility. Since not every concept has a definition on DBpedia, the hypothesis that the concept has no definition on DBpedia is not significant for the recommendation. We would explain it through a practical case.

Setting **Natural Language Processing** as an initial Q-keyword, Figure 3.7 shows the range of the directional SPARQL query and nodes that its concept importance would be calculated. By means of the proposed Concept importance estimation approach, we got the ranking list here:

Table 3.2: Top 10 concepts related to Natural Language Processing sort by Semantic-aware PageRank.

From	Concept	PageRank value
parent	Artificial intelligence applications	0.04195
parent	Natural language and computing	0.04156
siblingparent	Character encoding	0.03461
siblingparent	Computing by natural language	0.03171
siblingparent	Computational linguistics	0.03020
siblingparent	Language software	0.02873
grandparent	Linguistics	0.02483
siblingparent	Internationalization and localization	0.02327
child	Corpus linguistics	0.02227
siblingparent	Language-specific Linux distributions	0.02150

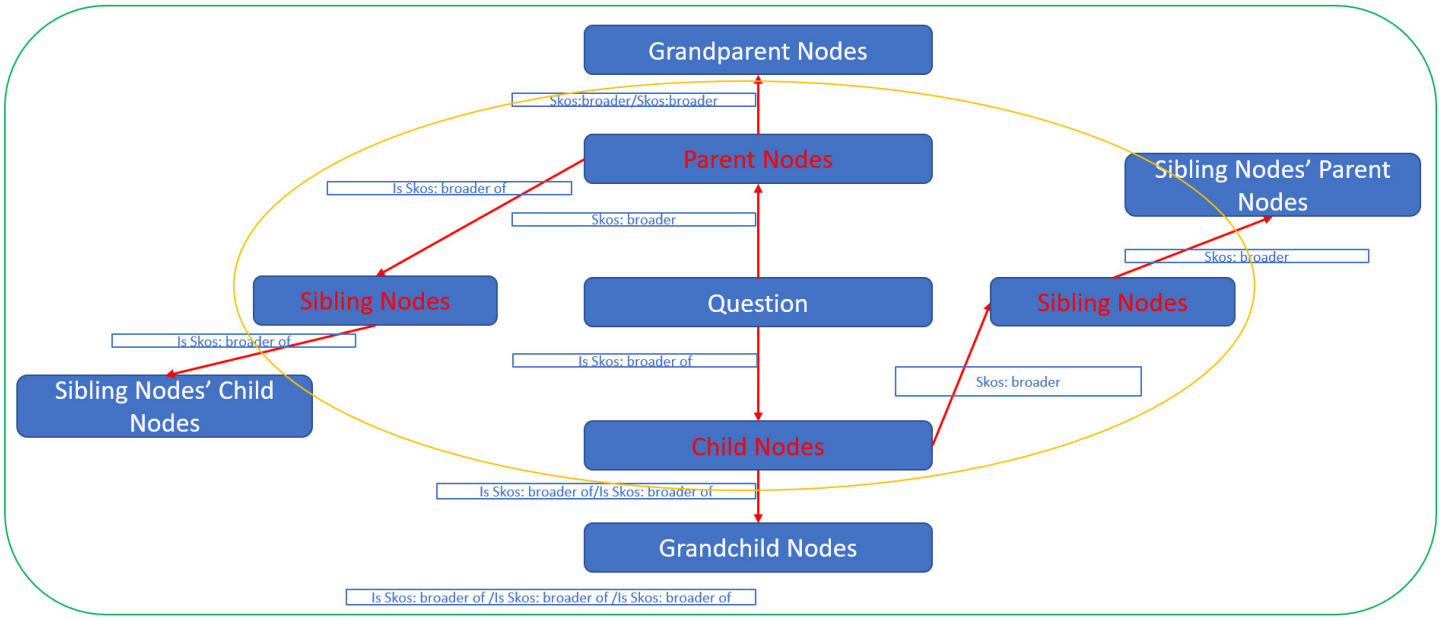

Figure 3.7: The range and nodes will be applied for the practical case.

By sending the SPARQL query to the public SPARQL endpoint as follow(Figure 3.8), we could easily discover that a concept of **Language-specific Linux distributions** has no definition on DBpedia, and it would be filtered.

```
SELECT DISTINCT ?definition

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

Figure 3.8: The SPARQL query for filtering.

Chapter 4

Evaluation

In this section, the ranking results would be analyzed by Spearman's correlation coefficient. Spearman's correlation coefficient measures the strength and direction of the association between two ranked variables. Furthermore, a case study would be conducted in order to test the hypothesis that using the Semantic-awareness recommendation with linked open data could help learners strengthen the knowledge construction process by discovering semantic related concepts during Web-based investigative learning.

4.1 Comparison between Semantic-aware PageRank, DBpagerank and User expectation

This experiment would be conducted to measure the strength of the association between Semantic-aware PageRank, DBpagerank and User expectation against the concepts extracted by the regulated concept map. DBpagerank[16] is the PageRank value for all the resources in DBpedia calculated by the DBpedia community. For the ranking list arranged by Professor, we consider it as the user expectation. In total three pairs of variables and the direction of the relationship would be analyzed. We would employ two initial Q-keywords, **Machine learning** and **Smoking**, and two regulated query strategy(More general and More specific) would be applied. Three pairs of variables would be analyzed:

Pair A

- Ranking list sort by Semantic-aware PageRank.
- Ranking list sort by DBpagerank.

Pair B

- Ranking list sort by DBpagerank.
- Ranking list arranged by Professor.

Pair C

- Ranking list sort by Semantic-aware PageRank.
- Ranking list arranged by Professor.

All of the recommended concepts and its ranking(If the values are same, the rank of the average value is returned.) are shown as follows:

Table 4.1: **More general** concepts of **Machine learning** recommended by proposed method and its ranking.

Concepts	Ranked by **PageRank value for RCM**	Ranked by **DBPageRank**	Ranked by **Professor**
Artificial_intelligence	1	4	1
Learning	2	12	2
Personhood	3	14	14
Cognition	4	11	12
Memory	5	9	16
Computational_neuroscience	6	10	3
Cybernetics	7	7	4
Unsolved_problems_in_computer_science	8	19.5	11
Futurology	9	15	8
Emerging_technologies	10	16	7
Education	11	1	13
Cognitive_science	12	5	6
Neuroscience	13	3	5
Formal_sciences	14	19.5	18
Behavior	15	13	15
Computer_science	16	2	9
Intelligence	17	8	10
Neuropsychological_assessment	18	17.5	20
Cognitive_neuroscience	19	6	17
Euthenics	20	17.5	19

Table 4.2: **More specific** concepts of **Machine learning** recommended by proposed method and its ranking.

Concepts	Ranked by **PageRank value for RCM**	Ranked by **DBPageRank**	Ranked by **Professor**
Evolutionary_algorithms	1	17	17
Artificial_neural_networks	2	12	9
Markov_models	3	17	12
Cluster_analysis	4	4	15
Graphical_models	5	10.5	13
Statistical_natural_language_processing	6	17	7
Genetic_algorithms	7	9	18
Dimension_reduction	8	17	3
Data_mining	9	1	14
Loss_functions	10	17	2
Bayesian_networks	11	17	10
Inductive_logic_programming	12.5	7	20
Computational_learning_theory	12.5	8	8
Gene_expression_programming	14	13	19
Algorithms	15	3	1
Deep_learning	16	2	5
Support_vector_machines	17	10.5	6
Markov_networks	18	17	11
Multivariate_statistics	19	5	16
Ensemble_learning	20	6	4

Table 4.3: **More general** concepts of **Smoking** recommended by proposed method and its ranking.

Concepts	Ranked by **PageRank value for RCM**	Ranked by **DBPageRank**	Ranked by **Professor**
Determinants_of_health	1	19	3
Addiction	2	6	1
Medical_ethics	3	8	2
Particulates	4	14	12
Social_inequality	5	16	11
Human_behavior	6	12	8
Fire	7	5	19
Habits	8	19	7
Mental_health	9	4	5
Health	10	3	9
Neuroscience	11	1	18
Environmental_health	12	10	10
Pollution	13	11	13
Behavior	14.5	15	15
Humans	14.5	17	17
Nervous_system	16.5	9	6
Branches_of_biology	16.5	19	16
Public_health	18.5	2	4
Natural_environment	18.5	7	14
Heat_transfer	20	13	20

Table 4.4: **More specific** concepts of **Smoking** recommended by proposed method and its ranking.

Concepts	Ranked by **PageRank value for RCM**	Ranked by **DBPageRank**	Ranked by **Professor**
Tobacco_companies	1	13	8
Tobacco	2	1	1
Cigarettes	3	7	14
Cigars	4	9	2
Tobacco_advertising	5	14	9
Tobacco_control	6	3	7
Cigar_brands	7	18	17
Smoking_cessation	8	5	3
Pipe_smoking	9	8	13
Electronic_cigarettes	10	18	15
Smoking_in_China	11.5	10	4
Smoking_in_Hong_Kong	11.5	18	6
Cannabis_smoking	13	4	18
History_of_tobacco	14	12	10
Tobacco_in_the_United_States	15	11	12
Tobacconists	16	18	11
Cigarette_holders	17	18	16
Cannabis	18	2	20
Smoking_in_India	19	15	5
Drugs	20	6	19

4.1.1 Analysis of Spearman's Correlation Coefficient

Spearman's correlation coefficient[7] is a statistical measure of the strength of a *monotonic* relationship between paired data. In a population, it is denoted by r_s and is by design constrained as follows:

$$-1 \leq r_s \leq 1$$

According to the definition, the closer r_s is to ± 1, the stronger the monotonic relationship. Since the correlation is an effect size, we could verbally describe the strength of the correlation using the following guide for the absolute value of r_s[7]:

- 0.00 - 0.19 : Very weak
- 0.20 - 0.39 : Weak
- 0.40 - 0.59 : Moderate
- 0.60 - 0.79 : Strong
- 0.80 - 1.0 : Very strong

For determining the significance of this test, we have to test the null hypothesis H_0 where there is no monotonic correlation the population ,against the alternative hypothesis H_1, where there is monotonic correlation. Let ρ_s as the Spearman's population correlation coefficient then we can thus express this test as follow:

$$H_0 : \rho_s = 0$$

$$H_1 : \rho_s \neq 0$$

$$\alpha = 0.05$$

4.1.2 Analysing Results

The analysing results for those three pairs of ranking lists are as follows:

Table 4.5: The analysing results of of Spearman's Correlation Coefficient.

Machine learning (more general)	Pair A	Pair B	Pair C
Coefficient :	0.0587	0.4304	**0.5714**
N :	20	20	20
T Statistic :	0.2494	2.0230	2.9542
Degree of Freedom :	18	18	18
P-value :	0.8058	0.0582	**0.0085***
Machine learning (more specific)	Pair A	Pair B	Pair C
Coefficient :	**-0.3560**	-0.0607	-0.2497
N :	20	20	20
T Statistic :	1.6161	0.2581	1.0941
Degree of Freedom :	18	18	18
P-value :	0.1235	0.7993	0.2883
Smoking (more general)	Pair A	Pair B	Pair C
Coefficient :	-0.0942	0.1325	**0.4938**
N :	20	20	20
T Statistic :	0.4016	0.5673	2.4092
Degree of Freedom :	18	18	18
P-value :	0.6927	0.5775	**0.0269***
Smoking (more specific)	Pair A	Pair B	Pair C
Coefficient :	0.1743	0.0682	**0.4573**
N :	20	20	20
T Statistc :	0.7510	0.2900	2.1817
Degree of Freedom :	18	18	18
P-value :	0.4623	0.7752	**0.0426***

***P<0.05, Two tailed**

By the observation of the analysing results above, for the recommendation list of **Machine learning(More general)**($r_s = 0.5714$, $n = 20$, $P < 0.05$), **Smoking(More general)**($r_s = 0.4938$, $n = 20$, $P < 0.05$) and **Smoking(More specific)**($r_s = 0.4573$, $n = 20$, $P < 0.05$), Pair C maintained the highest value of the Spearman's Correlation Coefficient. It shows that, in most cases, there are a **moderate** , **positive monotonic** correlation between Ranking lists sort by Semantic-aware PageRank value and ranking lists arranged by Professor(User expectation).

4.2 Case Study

The criteria of this case study are as follows: By applying two directional SPARQL query strategies(Figure 4.1) to 4 question keywords which are **Machine learning**, **Nuclear power**, **Governance**, and **Smoking**, we test the efficiency of the proposed recommendation system in Web-based investigative learning. By asking participants to pick out those concepts that are more general or more specific in meaning to the initial question keyword in the limited time. Meanwhile, the link to the concepts' definition page on DBpedia will be provided as a reference during the tasks. It simulates the navigational learning of Web-based investigative learning that learners could navigate those resources recommended by the proposed approach for knowledge construction process during the navigational learning stage.

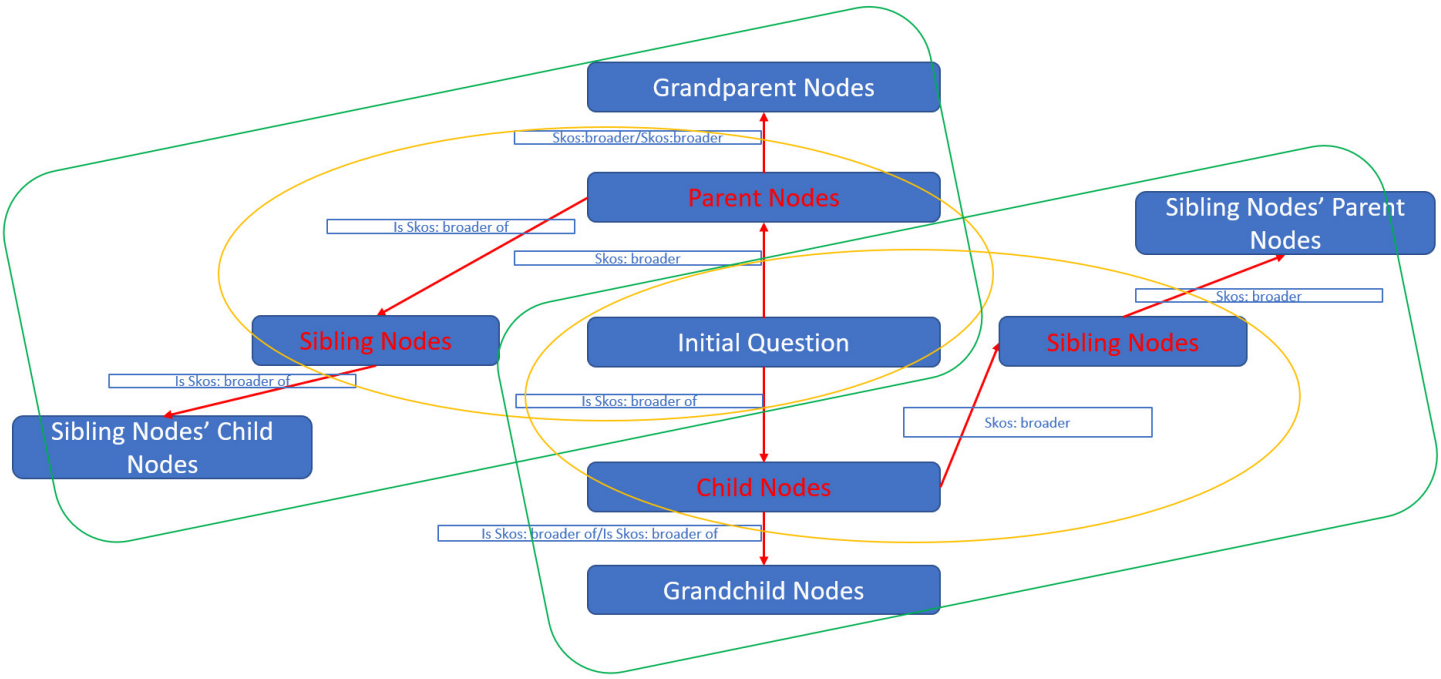

Figure 4.1: Two directional query strategies for the case study.

There are 4 sections in total. For each section, participants would process 3 tasks as follows:

- Picking out those concepts that are more general or more specific in meaning to the initial question keyword.

- Writing down three most important concepts they selected in the previous task.
- Post-test questionnaire.

In this case study, the **control variables** are as follows:

- Time is limited for each section(25 minutes for each, 100 minutes in total).
- Two directional query strategies would be included(Figure 4.1).
- 20 concepts would be shown as recommendations for each Q-keyword(10 for more general and 10 for more specific).
- Ranking list would be adjusted according to the requirement. For example, if the task is asking learner to pick out more general concepts to the initial Q-keyword, those top 10 concepts recommended by proposed method would be moved to the top.

Independent variable: Ordering of recommended concepts.

- Control group: Sort by DBpagerank.
- Experimental group: Sort by Semantic-aware PageRank.

Dependent variable(Comparison point):

- The number of concepts that participants pick out during the limited time.
- How well the important concepts determined by participants match the important concepts recommended by the system.

Target participants: 20 JAIST students.

4.2.1 Analyzing Results of First Task

For measuring the difference of the number of concepts that participants pick out during the limited time between two groups, the Mann-Whitney U Test[17] is employed. The purpose of this non-parametric measurement is to compare the difference between two populations. The basis on which we make inferences is also based on the sampling distribution composed of all the possible sample characteristics. Therefore, we test the hypothesis below:

- H_0 : There is no difference in number of concepts selected between two groups.
- H_1 : There is a difference in number of concepts selected between two groups.
- $\alpha = 0.05$
- Sampling distribution $Z_(critical) = \pm 1.96$

Instead of calculating the difference in the average, we calculate the verification statistical value U which is based on the grade of the variable score in the sample. For calculating the U value, we first merge all the observations from both groups to one set the two populations which are the number of concepts that participants pick out during the limited time, and then give the grades according to the value of the variable items. The higher value would maintain a higher grade, and then they would sort by order(from high to low). Then add up the grades assigned to each sample. Finally, compare the difference in the sum of levels between the two populations. Here we have:

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - \sum R_1$$

- N_1 and N_2 are the sample size of each group.
- $\sum R_1$ is the sum of the ranks in controlled group which is equal to 1305.

An equally valid formular U_2 is as follow:

$$U_2 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - \sum R_2$$

- $\sum R_2$ is the sum of the ranks in experimental group which is equal to 1935.

and

$$U = min(U_1, U_2) = 485$$

For large samples, U is approximately normally distributed. Therefore, the value of the standardized value $Z_(statistic)$ would be :

$$Z_(statistic) = \frac{U - \mu_u}{\sigma_u}$$

Where μ_u and σ_u are the mean and standard deviation. μ_u and σ_u are given by:

$$\mu_u = \frac{N_1 N_2}{2}$$

$$\sigma_u = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}$$

As a result,

$$Z_{(statistic)} = \frac{485 - 800}{73.94} = -4.26 < -1.960$$

Since $Z_{(statistic)} < -1.960$ with $\alpha = 0.05$(Two tailed), we have to reject H_0, and state that there is a difference in a number of concepts selected between the controlled group and experimental group.

Skewness is a measure of symmetry, and Kurtosis describes the tail shape of the data's distribution. By the observing Descriptive statistics(Table 4.6) for two groups in this case study, the data of the experimental group forms a negative skewness. The data distribution of the experimental group is left-skewed which means during the task of the experimental group, learners tend to select more concepts than the controlled group. Moreover, the data distribution of the experimental forms a positive kurtosis which indicates a fat-tailed distribution. It refers to an increase in the probability of concepts being selected for an extreme number in the experimental group.

Table 4.6: Descriptive statistics for two groups in this case study.

	Concetps selected in controlled group	Concetps selected in experimental group
Mean	7.025	**8.25**
Standard Error	0.4244	0.3257
Median	7	8
Mode	4	8
Standard Deviation	2.6841	2.0600
Sample Variance	7.2045	**4.2436**
Kurtosis	-0.4968	**1.7845**
Skewness	0.3070	**-0.9471**
Range	11	10
Minimum	3	3
Maximum	14	13
Sum	281	**330**
Count	40	40

4.2.2 Analyzing Results of Second Task

During the second task, participants were asked to write down the three most important concepts they selected in the previous task. In previous

task, there are 20 concepts in the recommendation list(10 for more general and 10 for more specific). The ranking list would be adjusted according to the requirement. For example, in the controlled group, if the task is asking the participants to pick out more general concepts to the initial Q-keyword, those top 10 concepts recommended by the proposed method sort by DBpagerank would be moved to the top. Similarly, in the experimental group, the top 10 concepts recommended by the proposed method sort by Semantic-aware PageRank, against the Regulated Concept Map would be moved to the top. Therefore, we slipped concepts in the recommendation list to the following three levels:

- Level 1 : Rank 1-3(sort by DBpagerank or Semantic-aware PageRank)
- Level 2 : Rank 4-10(sort by DBpagerank or Semantic-aware PageRank)
- Level 3 : Rank 11-20(sort by DBpagerank or Semantic-aware PageRank)

The observed results of the concepts selected by the participants corresponding to the levels are summarized in table 4.7. are summarized in the table The Chi-square test[22] of association evaluates relationships between those three levels above. We test the hypothesis as follow:

- H_0 : There is no relationship exists on the three levels in the population.
- H_1 : There is a relationship exists on the three levels in the population.
- $\alpha = 0.05$

Table 4.7: Concepts selected by the participants corresponding to the levels.

	Level 1	Level 2	Level 3	Row Totals
Controlled group	44	53	23	120
Experimental group	53	51	16	120
Column Totals	97	104	39	240

The calculation of the Chi-Square statistic is as follow:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

where f_0 is the observed counts in the cells and f_e is the expected frequency if NO relationship existed between three levels. By calculating, the chi-square

statistic is 2.1299. The p-value is 0.344741. The result is not significant at $p < 0.05$. We have to reject H_1, and state that there is no relationship exists on the three levels in the population. It means whatever the recommended concepts is sort by DBpagerank or Semantic-aware PageRank, participants still could pick out those concepts which are expected as important concepts by the proposed method under a certain probability.

4.2.3 Analyzing Results of Post-test Questionnaire

In the third task of each section, the post-test questionnaire was conducted. The purpose is to investigate the participants' perception of the satisfaction and effectiveness of the system. Likert Scale was employed as a measure. Regarding the satisfaction of the system, three questions will be asked:

- I am satisfied with the recommendations.
- The recommendations are useful to me.
- The recommendations are unanticipated.

To measure the difference of the participants' perception of the satisfaction and effectiveness of the system between the controlled group and experimental group, Mann-Whitney U Test[17] is employed. The purpose of this non-parametric measurement is to compare the difference between two populations. The basis on which we make inferences is also based on the sampling distribution composed of all the possible sample characteristics. Therefore, we test the hypothesis below:

- H_0 : There is no difference in the participants' perception of the satisfaction of the system between two groups.
- H_1 : There is a difference in the participants' perception of the satisfaction of the system between two groups.
- $\alpha = 0.05$
- Sampling distribution $Z_(critical) = \pm 1.96$

For calculating the U value, we first sum the scores of three questions and merge all the observations from both groups to one set. Then give the grades according to the value. The higher value would maintain higher grade, and then they would sort by order(from high to low). Then add up the grades

assigned to each sample. Finally, compare the difference in the sum of levels between the two populations. Here we have:

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - \sum R_1$$

- N_1 and N_2 are the sample size of each group.
- $\sum R_1$ is the sum of the ranks in controlled group which is equal to 1215.

An equally valid formular U_2 is as follow:

$$U_2 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - \sum R_2$$

- $\sum R_2$ is the sum of the ranks in experimental group which is equal to 2025.

and

$$U = min(U_1, U_2) = 395$$

For large samples, U is approximately normally distributed. Therefore, the value of the standardized value $Z_(statistic)$ would be :

$$Z_(statistic) = \frac{U - \mu_u}{\sigma_u}$$

Where μ_u and σ_u are the mean and standard deviation. μ_u and σ_u are given by:

$$\mu_u = \frac{N_1 N_2}{2}$$

$$\sigma_u = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1}{12}}$$

As a result,

$$Z_(statistic) = \frac{395 - 800}{73.94} = -5.478 < -1.960$$

Since $Z_(statistic) < -1.960$ with $\alpha = 0.05$(Two tailed), we have to reject H_0, and state that there is a difference in the participants' perception of the satisfaction of the system between two groups. By the observation of Descriptive statistics(Table 4.8) for the participants' perception of the satisfaction of the system between two groups, the data of both groups forms a negative skewness. The data distribution of both group are left-skewed

which means whatever the concepts recommended by proposed method is sort by DBpagerank or PageRank value against RCM, learners tend to be satisfied. However, only the data distribution of experimental forms a positive kurtosis which indicates a fat-tailed distribution. It refers to an increase in the probability of satisfaction scores being selected for an extreme number in experimental group.

Table 4.8: Descriptive statistics of the participants' perception of the satisfaction.

	Controlled group	Experimental group
Mean	11.675	**12.6**
Standard Error	0.2491	0.2449
Median	12	13
Mode	13	13
Standard Deviation	1.5752	1.5492
Sample Variance	2.4814	2.4000
Kurtosis	-0.2680	**1.2748**
Skewness	-0.7131	-0.7195
Range	6	7
Minimum	8	8
Maximum	14	15
Sum	467	504
Count	40	40

Similarly, regarding the effectiveness of the system, three questions will be asked:

- The recommendations are relevant to initial keyword.
- The recommendations enables me to strengthen the knowledge construction process.
- The recommendations makes investigation more efficient.

Therefore, we test the hypothesis below:

- H_0 : There is no difference in the participants' perception of the effectiveness of the system between two group.

- H_1 : There is a difference in the participants' perception of the effectiveness of the system between two group.
- $\alpha = 0.05$
- Sampling distribution $Z_(critical) = \pm 1.96$

For calculating the U value, we first sum the scores of three questions and merge all the observation from both groups to one set. Then give the grades according to the value. The higher value would maintain higher grade, and then they would sort by order(from high to low). Then add up the grades assigned to each sample. Finally, compare the difference in the sum of levels between the two populations. Here we have:

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - \sum R_1$$

- N_1 and N_2 are the sample size of each group.
- $\sum R_1$ is the sum of the ranks in controlled group which is equal to 1195.

An equally valid formular U_2 is as follow:

$$U_2 = N_1 N_2 + \frac{N_2(N_2+1)}{2} - \sum R_2$$

- $\sum R_2$ is the sum of the ranks in experimental group which is equal to 2045.

and

$$U = min(U_1, U_2) = 375$$

For large samples, U is approximately normally distributed. Therefore, the value of the standardized value $Z_(statistic)$ would be :

$$Z_(statistic) = \frac{U - \mu_u}{\sigma_u}$$

Where μ_u and σ_u are the mean and standard deviation. μ_u and σ_u are given by:

$$\mu_u = \frac{N_1 N_2}{2}$$

$$\sigma_u = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}$$

As a result,

$$Z_(statistic) = \frac{375 - 800}{73.94} = -5.748 < -1.960$$

Since $Z_(statistic) < -1.960$ with $\alpha = 0.05$(Two tailed), we have to reject H_0, and state that there is a difference in the participants' perception of the effectiveness of the system between the two groups. By the observation of Descriptive statistics(Table 4.9) for the participants' perception of the effectiveness of the system between the two groups, the data of both groups form a negative skewness. The data of the experimental group forms more serious negative skewness. The data distribution of both groups are left-skewed which means whatever the concepts recommended by proposed method is sort by DBpagerank or PageRank value against RCM, learners tend to be satisfied with the efficiency of the system. Moreover, the data distribution of experimental forms a positive kurtosis which indicates a fat-tailed distribution. It refers to an increase in the probability of the participants' perception of the effectiveness scores being selected for an extreme number in the experimental group.

Table 4.9: Descriptive statistics of the participants' perception of the effectiveness.

	Controlled group	Experimental group
Mean	12.5	**13.35**
Standard Error	0.2375	0.2280
Median	12.5	14
Mode	12	14
Standard Deviation	1.5021	1.4420
Sample Variance	2.2564	2.0795
Kurtosis	-0.2524	**1.5395**
Skewness	-0.2628	**-1.1959**
Range	6	6
Minimum	9	9
Maximum	15	15
Sum	500	534
Count	40	40

4.2.4 Discussion

By observing the analysis results above, three major issues affect the performance of the proposed method.

Firstly, there existed unexpected recommendations by the proposed method. Take the more general concepts of the initial question **Machine learning** recommended by the proposed method as an instance(Table 4.10). We could realize that those concepts such as **Neuropsychological assessment** and **Euthenics** are quite difficult for learner to construct the knowledge for the initial question even those concepts do exist semantic relationships between initial question over the web.

Table 4.10: More general concepts of Machine learning recommended by proposed method and its ranking.

Concepts	Ranked by **PageRank value for RCM**	Ranked by **DBPageRank**	Ranked by **Professor**
Artificial_intelligence	1	4	1
Learning	2	12	2
Personhood	3	14	14
Cognition	4	11	12
Memory	5	9	16
Computational_neuroscience	6	10	3
Cybernetics	7	7	4
Unsolved_problems_in_computer_science	8	19.5	11
Futurology	9	15	8
Emerging_technologies	10	16	7
Education	11	1	13
Cognitive_science	12	5	6
Neuroscience	13	3	5
Formal_sciences	14	19.5	18
Behavior	15	13	15
Computer_science	16	2	9
Intelligence	17	8	10
Neuropsychological_assessment	18	17.5	20
Cognitive_neuroscience	19	6	17
Euthenics	20	17.5	19

Secondly, we employed the semantic relations over the web to represent the interlinking between concepts extracted by the proposed method. However, as we mentioned in the section 2.1, making decision for interlinking between concepts is one of the knowledge management activities that learner would build an individual's cognitive network. During this process, learner may not to consider the semantic relations over the web. First task of the case study simulates the navigational learning of Web-based investigative learning that learners could navigate those resources recommended by the proposed approach for the knowledge construction process. Take Figure 4.2 as an instance, it could not be wrong when we consider this decision making as a cognitive process.

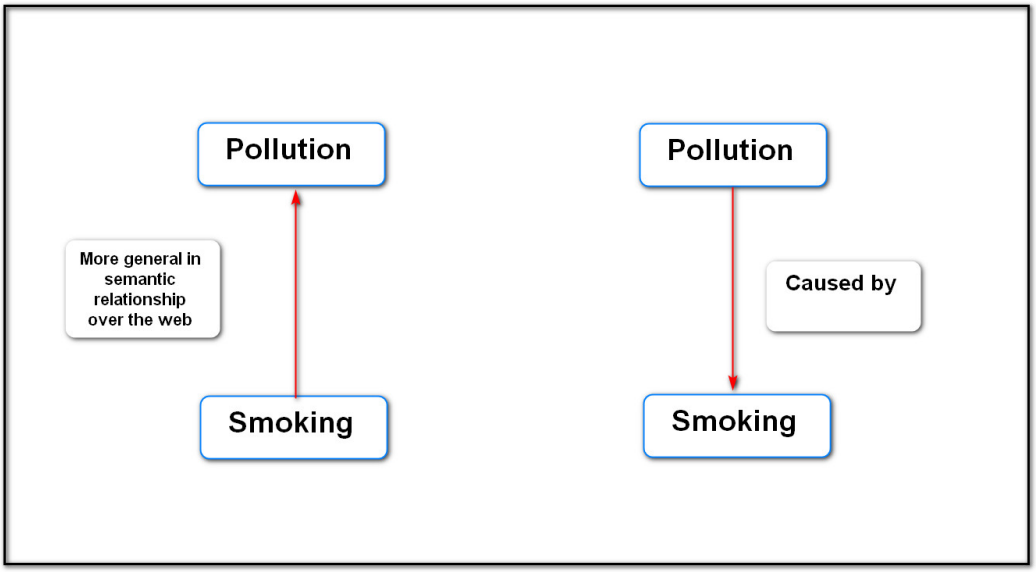

Figure 4.2: An example of interlinking between two concepts.

Thirdly, according to the concept importance estimation we proposed above, we assume that the importance(Semantic-aware PageRank) of a concept node in Regulated Concept Map is determined by the number of outbound links on that concept. It means the evaluation of the relevance[23] between the recommended concepts and initial Q-keyword was regardless. The analyzed results of the second task in the case study confirmed that whatever the recommended concepts is sort by DBpagerank or Semantic-aware PageRank, participants still could pick out those concepts which are expected as important concepts by the proposed method under a certain probability. However, there is a gap between the priority of the relevance and importance when participants making decision to pick out the related concept. For example, there existed a concept contain a serious Semantic-aware PageRank value over the Regulated Concept Map, but it is far a way from the initial Q-keyword in DBpedia. Therefore, participants may tend to mark this concept as the related but not important one.

Chapter 5

Conclusion

In this work, we propose a Semantic-awareness recommendation method of extracting and presenting related concepts at different levels for an initial question through LOD to promote the efficiency in the knowledge construction process in Web-based investigative learning. To prevent learners from the self-directed investigation, we propose the regulated concept map generation to retrieve the relevant concepts at different levels with LOD and Semantic relations. For evaluating the relevance between initial Q-keyword and concepts in the regulated concept map, we defined the relativity as Semantic relations, node importance, and content containment for concepts we extracted from DBpedia. Owing to the finding of analysing results measure by Spearman's Correlation Coefficient, we proposed the methodology of concept importance estimation maintained most serious strength of the association between learner's expectation. We have also reported a case study whose purpose was to evaluate that using the Semantic-awareness recommendation with linked open data could help learners strengthen the knowledge construction process by discovering semantic related concepts during Web-based investigative learning. The results of the study suggest that Semantic-awareness recommendation with linked open data promotes the efficiency of the knowledge construction process.

5.1 Future Work

First of all, the proposed method in this work only supports recommending concepts based on one initial question. In order to support long-term learning scenario creation, a recommendation against multiple Q-keywords is needed. Secondly, as we mentioned above, the capabilities of LOD was underutilized. We could not tell that the semantic relations over the web are the best way

for recommending concepts. Therefore, by combining different relations over the web, the results could be highly-anticipated. Thirdly, there still is a room to improve the concept importance estimation and filtering, such as applying certain techniques like machine learning and natural language processing.

Bibliography

[1] DBpedia Blog: new 2016-04 dbpedia release. https://blog.dbpedia.org/2016/10/19/yeah-we-did-it-again-new-2016-04-dbpedia-release/. Accessed: 2021-1-18.

[2] LOD cloud diagram: the linked open data cloud. https://lod-cloud.net/. Accessed: 2021-1-27.

[3] Tim Berners-Lee design issues: Linked data. https://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2021-1-18.

[4] W3C Recommendation: resource description framework (rdf) model and syntax specification. https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/. Accessed: 2021-1-18.

[5] W3C Recommendation skos simple knowledge organization system reference. https://www.w3.org/TR/skos-reference. Accessed: 2021-1-18.

[6] W3C Recommendation: sparql protocol for rdf. https://www.w3.org/TR/rdf-sparql-protocol/. Accessed: 2021-1-18.

[7] *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.

[8] Tim Berners-Lee and CERN. Information management: A proposal. 1989.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30:107–117, 1998.

[10] Marco de Gemmis Cataldo Musto, Giovanni Semeraro and Pasquale Lops. Tuning personalized pagerank for semantics-aware recommendations based on linked open data. *European Semantic Web Conference 2017*, pages 169–183, 2017.

[11] H. H. Fu, K. J. Lin, and H. T. Tsai. Damping factor in google page ranking. *APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY*, 22:431–444, 2005.

[12] M. Hagiwara, A. Kashihara, S. Hasegawa, K. Ota, and R. Takaoka. Adaptive recommendation for question decomposition in web-based investigative learning. In *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*, pages 1–9, 2019.

[13] Shiori Ichinose, Ichiro Kobayashi, Michiaki Iwazume, and Kouji Tanaka. Ranking the results of dbpedia retrieval with sparql query. In *Revised Selected Papers of the Third Joint International Conference on Semantic Technology - Volume 8388*, JIST 2013, page 306–319, Berlin, Heidelberg, 2013. Springer-Verlag.

[14] Akihiro Kashihara and Naoto Akiyama. Learner-created scenario for investigative learning with web resources. In H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik, editors, *Artificial Intelligence in Education*, pages 700–703, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[15] K. Kashihara and N. Akiyama. A model of meta-learning for web-based navigational learning. *International J. of Advanced Technology for Learning*, 2(4):198–206, 2005.

[16] J. Lehmann, R. Isele, M. Jentzsch, and et al. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[17] Patrick E. McKnight and Julius Najab. *Mann-Whitney U Test*, pages 1–1. American Cancer Society, 2010.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*, SIDL-WP-1999-0120(1999-66), 1999.

[19] G. Siemens. Connectivism: A learning theory for the digital age. *elearnspace.org*, 2012.

[20] Paolo Tomeo Tommaso Di Noia, Vito Claudio Ostuni and Eugenio Di Sciascio. Sprank: Semantic path-based ranking for top-n recommendations using linked open data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(9):1–34, 2016.

[21] D. Vrandečić and M. Kroetzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[22] Karl L. Wuensch. *Chi-Square Tests*, pages 252–253. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[23] T. Yamauchi. Relevance between q-keywords corresponding to transition of interest in web-based investigative learning. Master's thesis, Japan Advanced Institute of Science and Technology, School of Information Science, Nomishi Ishigawa, 2020.

Publication

1. Kang TING and Shinobu HASEGAWA: Semantic-awareness Recommendation with Linked Open Data in Web-based Investigative Learning. International Conference on Education and Distance Learning(2021 under submission).