Multi-Promoting Search Engine Re-Ranking on Domain-Specific Knowledge Representation

By
Yu Hou, BE, SE

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

at

School of Computer Science and Information Systems

Pace University

April 2020

We hereby certify that this dissertation, submitted by Yu Hou, satisfies the dissertation requirements for the degree of *Doctor of Philosophy in Computer Science* and has been approved.

____-________________

Dr. Lixin Tao Date
Chairperson of Dissertation Committee

____-________________

Dr. Li-Chiou Chen Date
Dissertation Committee Member

____-________________

Dr. Juan Shan Date
Dissertation Committee Member

School of Computer Science and Information Systems
Pace University 2020

Abstract

Multi-Promoting Search Engine Re-Ranking on Domain-Specific Knowledge Representation

By
Yu Hou, BE, SE

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

April 2020

As the tsunami of data has emerged, search engines have become the most powerful tool for obtaining scattered information on the internet. It is designed to search for data stored on a computer system or the World Wide Web. The ranking algorithm plays a vital role in determining the relevance and weight of the results to search queries (or keywords). The traditional search engines- the most extensively utilized search methods, has several ranking methods, such as Keyword-based web pages ranking and link analysis (PageRank algorithm and HITS algorithm). However, these algorithms must combine the keyword-based algorithm to optimize the ranking of web pages. This leads to the lack of ability of the traditional search engine in semantic understanding. People expect that the search engines could understand their searching inquiry by content meanings rather than literal strings.

The semantic search engine can execute the extensive and effective semantic reasoning in the network environment by annotating the resource objects in the network. However, annotating every web page leads to inefficiency for the search engine. This dissertation designed a multi-promoting approach to alleviate this problem, and the approach could improve the current traditional keyword-based search, and emulate the effects of semantic search. The first method is to improve the existing solution by introducing the ontology as the domain knowledge. The semantic distance, semantic coincidence, and level difference are used to calculate the semantic similarity among the concepts in the ontology. Then, the candidate documents can be ranked based on the semantic similarity and the TF-IDF weight. The second method is to improve the existing solution by introducing the knowledge graph as domain knowledge. Entities and relationships within the knowledge graph are embedded, and the cosine similarity is used to calculate the semantic similarity among the embedded entities in the knowledge graph. Then, the candidate documents are ranked according to semantic similarity and TF-IDF weights. Two methods work in the different knowledge bases ontology and knowledge graph, respectively. As knowledge expression becomes richer, the ability of search engines to identify and retrieve semantic-related concepts grows as well, and the performance of the search engine can be improved.

The experimental results and the response from the experimental questionnaire conducted by experts proved this.

The proposed methods are expected to achieve two desired outcomes, first, being able to rank the related semantic document to the top; second, being able to collect the related semantic document even they do not contain the query keyword. The experiment used academic articles in the field of Education Sciences from ERIC (https://eric.ed.gov/) as the experimental data set and used ‘exploration’- one of the concepts in ‘cognitive presence’ as a keyword to retrieve academic articles in ERIC. From the experimental questionnaire conducted by experts, we confirmed the experimental hypothesis and proved the methods proposed in this dissertation can improve the current traditional keyword-based search and emulate the effects of semantic search. This dissertation designed a recommendation system to optimize the user experience and improve the performance of domain knowledge in search engine. The recommendation system has been designed based on the introduced domain knowledge to help users narrow down the search scope during the retrieval process.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Lixin Tao for the continuous support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank the rest of my dissertation committee: Prof. Li-Chiou Chen and Prof. Juan Shan, for their insightful comments and encouragement.

Last but not least, I would like to thank my family: my parents, my wife, and my lovely son for supporting me spiritually throughout writing this dissertation and my life in general.

Table of Contents

Abstract iii

Table of Figures ix

Table of Tables xiii

Chapter 1 Introduction 1

1.1 Background 1

1.2 Motivation 2

1.3 Traditional Search Engines 4

1.4 Semantic Web Search Engine 10

1.5 Research Challenges 19

1.6 Contributions 21

1.7 The list of sub-tasks 22

1.8 Road Map 23

Chapter 2 Related Work 24

2.1 Knowledge Representation 24

2.2 Ontology 26

2.3 Protégé 29

2.4 Apache Jena 31

2.5 Semantic Similarity 32

2.6 TF-IDF 33

2.7 Knowledge Graph 37

2.8 Pace Protégé 39

2.9 How a traditional Search Engine works 42

2.9.1 Index table construction 42

2.9.2 The query procedure 44

2.10 How a traditional semantic search engine works ... 45

2.11 Conclusion ... 49

Chapter 3 Methodologies ... 51

3.1 Ontology-based Re-ranking approach ... 51

3.1.1 The Workflow ... 51

3.1.2 The semantic similarity calculation ... 53

3.2 Knowledge Graph-based Re-ranking model ... 57

3.2.1 The Workflow ... 58

3.2.2 Relationship in the embedding space ... 60

3.2.3 Knowledge embedding by Translating on Hyperplanes ... 62

3.2.4 Semantic Similarity ... 66

3.3 The candidate documents re-ranking ... 68

3.4 Recommendation system based on the domain knowledge ... 71

3.5 Conclusion ... 74

Chapter 4 Experiment ... 76

4.1 Introduction ... 76

4.2 Experiment Design ... 78

4.2.1 Data Preparation ... 80

4.2.2 The domain knowledge ... 82

4.2.3 A prototype for re-ranking models ... 86

4.2.4 The related work for the prototype design ... 87

4.3 The details of the re-ranking prototyping ... 89

4.4 Query Testing ... 103

4.5 Experiment Evaluation ... 106

4.6 Questionnaire results ... 108

Chapter 5 Conclusion ... 112

References 115

Appendix A Retrieval Ranking Questionnaire 120

Table of Figures

Figure 1 The example of the limitation in traditional search engine 3

Figure 2 The example of the basic idea of the Page Rank algorithm 8

Figure 3 Authority score and Hub score calculation 9

Figure 4 An example of HITS 10

Figure 5 An example of Google search 13

Figure 6 An example of Linked Data 14

Figure 7 An RDF graph representing Eric Miller. 15

Figure 8 A chunk of RDF in RDF/XML 16

Figure 9 Hierarchical structure of semantic web 17

Figure 10 An example of knowledge representation 25

Figure 11 The illustration of Taxonomy 29

Figure 12 Demo of the Protege user interface 30

Figure 13 A demonstration of IDF 34

Figure 14 A demonstration of ontology 38

Figure 15 Demonstration of a knowledge graph 39

Figure 16 Pace Protege GUI Interface with Custom Relations Features........................ 42

Figure 17 The initial index table... 43

Figure 18 Final index table 44

Figure 19 Search engine return the results for the query word$_2$ 45

Figure 20 SLR as query in the index table.. 46

Figure 21 A domain knowledge of SLR 47

Figure 22 Add an annotation on a webpage... 48

Figure 23 The index table with www.goodPhoto.com 49

Figure 24 The Framework of the Ontology-based re-ranking approach 53

Figure 25 Example of the Semantic Distance... 54

Figure 26 The example of semantic coincidence.. 55

Figure 27 The example of the Level Difference... 56

Figure 28 The framework of the knowledge graph-based model 59

Figure 29 The illustration of TransE.. 60

Figure 30 The illustration of TransH 64

Figure 31 Cosine Similarity.. 67

Figure 32 A demonstration of an article 71

Figure 33 The workflow of the recommendation system 72

Figure 34 Demonstration of sub-domain knowledge from the recommendation system. 74

Figure 35 The overall architecture of the first knowledge base....................................... 84

Figure 36 The overall architecture of the second knowledge base.................................. 85

Figure 37 The workflow of the prototype... 87

Figure 38 The workflow of the re-ranking prototype 90

Figure 39 Query in ERIC 91

Figure 40 The query results 92

Figure 41 The exported result from ERIC (a)... 93

Figure 42 The exported result from ERIC (b) 93

Figure 43 The Creation of a Django project via PyCharm 96

Figure 44 The registration of the application in setting.py... 97

Figure 45 Configure the view.py 97

Figure 46 Configure the url.py... 98

Figure 47 Start the Django server 98

Figure 48 The selected options in prototype’s interface. 99

Figure 49 The prototype's interface 99

Figure 50 The summary information of knowledge base in re-rank mode 1................. 100

Figure 51 The summary information of knowledge base in re-rank mode 2................. 101

Figure 52 The design of the search page with recommendation system 102

Figure 53 The search results in the page with recommendation system......................... 103

Table of Tables

Table 1 The example in the TF-IDF calculation 35

Table 2 Examples of matched documents by different models. Bold text is the related concept. 104

Table 3 The results review for the questionnaire 109

Chapter 1 Introduction

This dissertation designed a multi-promoting approach to improve the current traditional keyword-based search and emulate the effects of semantic search. It detailed two methods (ontology-based model and knowledge graph-based model) to convert the knowledge into the quantitative results and calculate the semantic similarity among the knowledge based on the quantitative results. As a result, the search engine can achieve the capabilities to capture the conceptualizations involved in users' intention and documents content, meaning, ultimately, the users' search experience can be improved.

1.1 Background

In 2014, there were 2.4 billion Internet users. That number grew to 3.4 billion by 2016, and in 2017, 300 million internet users added – making a total of 3.8 billion internet users in 2017. This situation was a 42% increase in people using the Internet within three years. Almost ninety percent of data in the world today had been created in the last two years. The current output of data on the global Internet is roughly 2.5 quintillion bytes a day. Domo, a company specializes in business intelligence tools and data visualization, has released its fifth annual infographic looking into the world's data generation and online behavior, called Data Never Sleeps 5.0. The Data Never Sleeps 5.0 shows that since 2013, the number of Tweets each minute has increased by 58% to more than 45,000 Tweets per minute in 2017. YouTube usage more than tripled from 2014-2016, with users uploading

400 hours of new video each minute of every day. Moreover, in 2017, users are watching 4,146,600 videos every minute. At the same time, 3,607,080 Google searchers conducted worldwide each minute of every day. All these numbers signify that a tsunami of data has emerged, which is becoming increasingly important information for internet users. Although the big data could help users to capture useful information, however, it is hard for the users to search the satisfying information when they are facing a vast amount of data. Thus, the search engine has become the most powerful tool for obtaining scattered information on the Internet.

1.2 Motivation

As the amount of data increases rapidly in our daily life, it causes many challenges for the search engine. When the user submits a query, traditional search engines only return results that match the query, but the information may not match the user's real need. Due to some apparent limitations of conventional search engines (such as the search results can only retrieve with the query keywords), Query-related semantics can easily be ignored by traditional search engines. For example, if a user wants to search "iPhone" in a search engine. Although a web page mentioned, "Apple is trying to make a new type of mobile phone." Traditional search engines are powerless in this situation because they based on literal matching as the basis for sorting. The results with no literal matching (no keyword "iPhone" on the web page) will not be listed, though the query and the web page are semantically relevant.

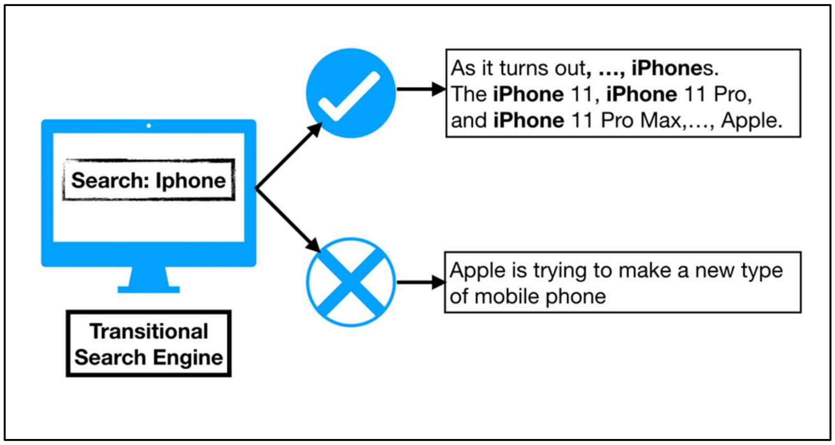

Search: Iphone
Transitional Search Engine
As it turns out, ..., iPhones. The iPhone 11, iPhone 11 Pro, and iPhone 11 Pro Max,..., Apple.
Apple is trying to make a new type of mobile phone

Figure 1 The example of the limitation in traditional search engine

One of the essential research aspects in search engines is "intention." It is generally defined as search the query and explores the semantic information behind the query. This study means the search engines expected to have the capabilities to understand users' searching intention by the content meanings rather than the literal strings.

Sometimes, in terms of traditional search engines, users always intend to add more keywords to increase search accuracy to narrow down the search. However, expecting all the users to be experts in any field is impossible. Therefore, this dissertation enhanced the traditional query methods in search engines, introduced domain knowledge in a particular area so that the search engine can become an expert in the field. By using the proposed methods in this dissertation, the search engine can help users get more accurate and comprehensive information in the field.

Most of the search platforms, such as Google Scholar, they embed a cross-domain nature, which not only requires the query methods to dig into domain topics piece by piece, but also requires the users to filter the related articles manually. Over the years, with increasing query needs to retrieve accurate information in some specific areas, people's attention has shifted from extensive searches to specific vertical domains, and some domain-specific searches have emerged.

As the introduction before, traditional search engines can only use query words, phrases, or sentences for retrieval. To get more relevant information, the user needs to perform multiple searches by using different query strings. Furthermore, the user may or may not know the exact relationship between the information (s)he collects.

To alleviate the limitations of the traditional query methods in search engines, the proposed methods in this dissertation could let the search engines have the capabilities to capture the conceptualizations involved in users' intention and web pages' content meaning. In short, when search engines obtain the semantic similarity among each concept in domain knowledge, they can know how relevant between every pair of concepts. When the users try to search a keyword, which is one of the concepts in the domain knowledge, the search engine will return the query results not only relate to the query keyword, but also its relevant concepts in the domain knowledge.

1.3 Traditional Search Engines

A search engine as a system designed to search for information stored on a computer system or the World Wide Web. The search results generally presented as a list. The search results presented to the users in a specific order, and a ranking algorithm generates this order. In general, a search engine uses the algorithm to determine the relevance and weight of the results to search queries (or keywords). Thus, the search engine will put the most relevant result in the first position, and so on. Many ranking methods have currently used worldwide. First, the traditional search engines, the two most popular web pages ranking methods in the traditional search engine, are keyword-based web pages ranking and link analysis.

The main idea of the keyword-based web pages ranking is to utilize the frequency and the position of the keyword in the web pages or documents for ranking the results in a search engine. The keyword-based ranking is one of the most mature ranking methods. It has been widely as the core technology of many search engines. The fundamental principle is that the higher frequency of the keyword in a web page or a document appears, the more critical the position is. Then, the better the correlation with the search term is considered.

A keyword frequency refers to how often a keyword appears on a given webpage or within a piece of content. The more frequently a keyword appears in a given page or part of the content, the higher the keyword frequency would be. A keyword position also refers to position weighting. In the traditional search engine, the keyword position mainly weights for the webpages. The different position of the keyword will get different weights, and the search engine will return the different results based on the weights. In general, the information about the keyword position that can be considered, such as whether the keyword located in a title, whether the keyword located in a body, the font size of the keyword, the keyword is bold or not.

As the introduction previously, keyword-based web pages ranking is still used by many search engines because it is easy to implement and credible. However, when a keyword contains a common word, such as “and,” “is.” Then the ability of this technology to judge the relevance is significantly weakened. So, for a massive number of web pages, their quality is uneven. Some low-quality web pages with a high frequency of certain words cause no valuable information in those web pages. Therefore, if we only consider the keyword-based relevance in the web pages ranking and do not consider the quality of the web pages themselves, so it is difficult to filter out the web pages without the vital information.

Link analysis refers to the multidimensional analysis of hyperlinks in a web structure. At present, the applications of the link analysis mainly reflected in network information retrieval, network metrology, data mining, and web structure modeling as one of Google's core technologies; link analysis algorithm applications have shown great value in many aspects. The goal of a link analysis ranking algorithm consists in inferring the importance of a web page based on the topological structure of the graph of the World Wide Web (WWW). The link analysis ranking algorithm runs through the web graph and analyzes the outgoing arcs and the incoming arcs of the pages. Therefore, each page of the site is associated with a value based on which the order placed. More specifically, we can simulate a web page on the Internet as a node, and the "out chain" of this web page regarded as a "directed edge" pointing to other nodes.

In contrast, the "incoming chain" is the directed edge of other nodes pointing to this node. Then the entire network becomes a directed graph by using this method. Moreover, the evaluation of web page quality follows two assumptions as below: 1. Quantity assumption: the more significant the ingress (number of links) of a node (web page), the higher the page quality is; 2. Quality assumption: the source of the ingress of a node (which pages are linking to it) has a higher quality, then the web page has a more top quality. The most recent and most refined algorithms are PageRank and HITS.

In 1996, Larry Page and Sergey Brin developed the PageRank at Stanford University. In 1998, the first paper was published; it manly describes the PageRank algorithm and the initial prototype of the Google search engine. [1]The PageRank algorithm assigns the same score to each page initially, then updating the PageRank score for each page by iterative, recursive calculation until the score is stable. More specifically, each page distributes its current PageRank score evenly to the "outgoing chains" (links to other web pages); thus, each link obtains the corresponding weights. Moreover, the pages' new PageRank score is equal to the default score plus the weights passed from the links. Figure 2 shows the basic concept of PageRank. If there are many links link to one web page (such as B in the figure), we can consider that web page B has high quality. If there is a high-quality web page link to another one (such as B link to C), then the linked web page C can be considered as a high-quality web page as well.

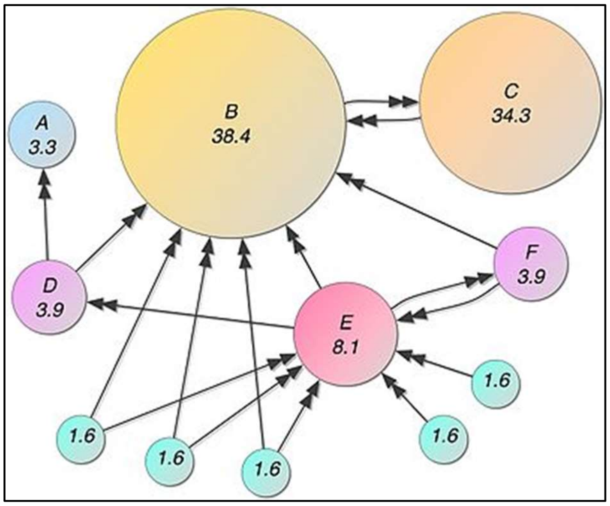


Figure 2 The example of the basic idea of the Page Rank algorithm

The HITS algorithm [2] is also fundamental and essential in link analysis, and the Teoma search engine has implemented it. Two basic definitions in the HITS algorithm are "Authority page" and "Hub page." The "Authority" page refers to high-quality web pages related to a particular field or a topic, such as Google's homepage is a high-quality web page in the search engine field. The "Hub" page refers to a web page that contains many links to high-quality "Authority" pages. The goal of the HITS algorithm is to find high-quality "Authority" pages and "Hub" pages related to the users' queries in the massive web pages. There are two assumptions in the HITS algorithm: 1. Many "Hub" pages will point to an "Authority" page; 2. A "Hub" page will lead to a lot of "Authority" pages. Through these two underlying assumptions, the mutual enhancement relationship between the Hub page and the Authority page can be derived. That is, the linked page will get higher quality when it has a higher quality of a page's Hub. The reverse is also true. The pages will have a higher quality link to them, and then they will have a higher quality of authority. By continuously iterating through this mutually reinforcing relationship, we can find out which pages are high-quality Hub pages and which pages are high-quality Authority pages. The authority score can be defined as:

$$A_p = \sum_{q:q \rightarrow p} H_q \tag{1}$$

Moreover, the hub score can be defined as:

$$H_p = \sum_{q:p \rightarrow q} H_q \tag{2}$$

Figure 3 shows an example of the authority score and hub score calculation.

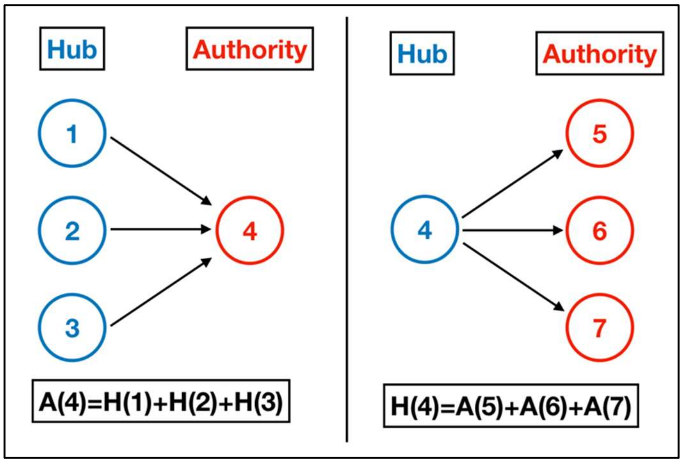


Figure 3 Authority score and Hub score calculation

When we obtained the authority score and the hub score, then every page must be given initial scores, and final scores are computed by successively repeating the summing processes with normalization until a predefined criterion is satisfied. Figure 4 shows an example of the HITS algorithm.

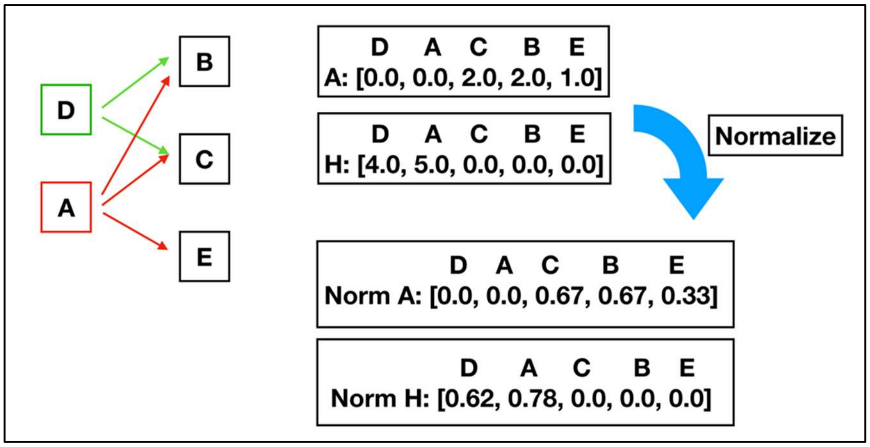

D
A
B
C
E
D A C B E
A: [0.0, 0.0, 2.0, 2.0, 1.0]
D A C B E
H: [4.0, 5.0, 0.0, 0.0, 0.0]
Normalize
D A C B E
Norm A: [0.0, 0.0, 0.67, 0.67, 0.33]
D A C B E
Norm H: [0.62, 0.78, 0.0, 0.0, 0.0]

Figure 4 An example of HITS

It should be noted that the PageRank algorithm is topic independent. Namely, it is not related to the query input by the user. A web page with a high PageRank value only can be considered as an essential web page. Which means, the PageRank algorithm should be used in combination with the content correlation algorithm. For the HITS algorithm, before the algorithm iteratively calculates the page of the highest-ranking authority for a specific search question, the query needs to be submitted to an existing search engine (or its own constructed retrieval system) and returned. In the search results, the top-ranked webpage is extracted, and a set of initial webpages highly correlated with the user query is obtained, and the collection is called a root set. Then, running the HITS algorithm on the root set can get the best results.

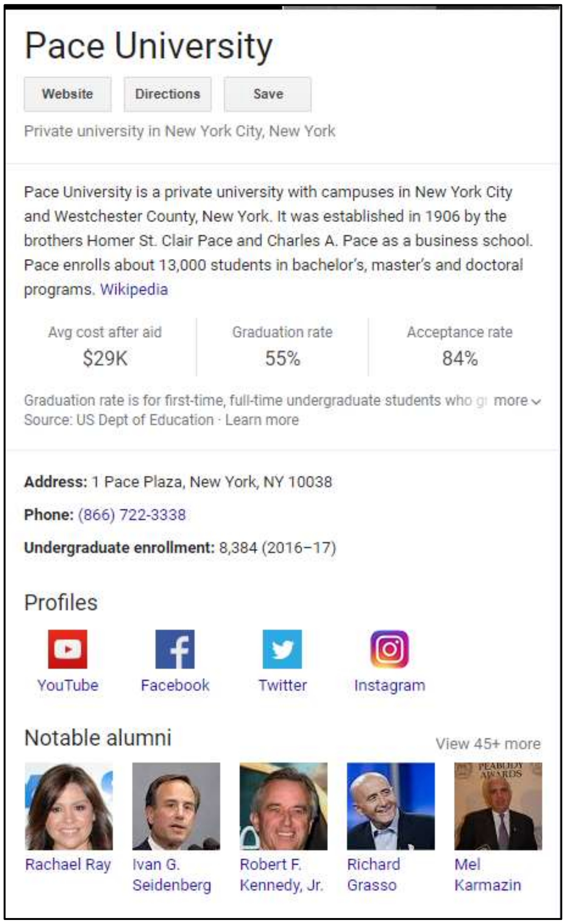
1.4 Semantic Web Search Engine

It is not difficult to see that all the algorithms listed above need to combine the keyword-based algorithm to optimize the ranking of web pages. In other words, the search engines return the ranked results to the users only consider the maximum number of the keyword occur in the web pages. Therefore, traditional keyword-based search engines bear several limitations. For example, the traditional keyword-based search engines are hard to be recognized the abbreviation and terms which are similar to the keyword. Even the traditional keyword-based search engines could increase the importance of this web page. For example, if a user wants to search "iPhone" in a search engine, although a web page mentioned about "Apple is trying to make a new type of mobile phone," traditional search engines are powerless in this situation because they based on literal matching as the basis for sorting. The result of no literal matching (no keyword "iPhone" on the web page) will not be searched, the query and the web page are semantically relevant. Even the most widely used search engines return useless pages to the users commonly. Therefore, to improve users' search experience and assist them in achieving more useful and accurate result has become an essential challenge for search engines.

The third generation of the search engine has emerged to solve the problems in the keyword-based web pages ranking. The Semantic Web is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C). The term was proposed by Berners-Lee, [3] and he described the semantic web as a component of "Web 3.0". It describes the development of the web, which consists of human-readable documents, computer-controlled data, and information. The Semantic Web is an operational information network, that is, information used to interpret symbols derived from data through the semantic theory. The semantic theory describes the meaning, in which the logical connection of terms establishes interoperability between systems. Tim

Berners Lee described this at the first World Wide Web conference in 1994. However, this simple idea has not been realized to a large extent.

In recent years, the web has evolved from the initial linking of human-readable documents to the linking of documents and data, such as Google. However, Google only helps users to find the relevant entities in the knowledge graph. For example, when we search pace university (see figure 5), Google will return the related entity with the pace university in the knowledge graph, such as an address, phone number, famous alumni. This result is not enough to help us find the most relevant information from a large number of web pages or documents. Therefore, this research can enable search engines to retrieve web pages or document more semantically based on the relevance between the query and web pages or document content.



Pace University

Private university in New York City, New York

Pace University is a private university with campuses in New York City and Westchester County, New York. It was established in 1906 by the brothers Homer St. Clair Pace and Charles A. Pace as a business school. Pace enrolls about 13,000 students in bachelor's, master's and doctoral programs. Wikipedia

Avg cost after aid	Graduation rate	Acceptance rate
$29K	55%	84%

Graduation rate is for first-time, full-time undergraduate students who g more
Source: US Dept of Education · Learn more

Address: 1 Pace Plaza, New York, NY 10038

Phone: (866) 722-3338

Undergraduate enrollment: 8,384 (2016–17)

Profiles

Notable alumni

Figure 5 An example of Google search

Supporting this development is a set of technologies for linking structured data published on the web, called linked data. Linked Data refers that by using the web to create typed links between data from different sources. [4] Figure 6 shows part of the Linking Open Data (LOP) project cloud diagram. Berners-Lee [5] outlined a set of rules for publishing data on the web in a way that all published data becomes part of a single global data space: 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

Data relies on two technologies that are fundamental to the Web: Uniform Resource Identifiers (URIs) [6] and the HyperText Transfer Protocol (HTTP). [7]

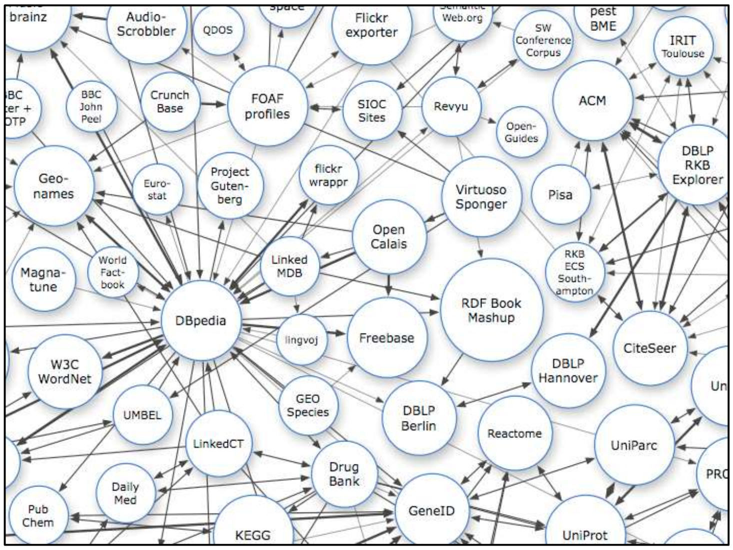


Figure 6 An example of Linked Data

Consistent with the need for a Web semantics, the user community, including standards organizations like the Internet Engineering Task Force and the World Wide Web Consortium (W3C), has directed significant efforts at specifying, developing, and deploying languages for sharing meaning. These languages provide a foundation for semantic interoperability. The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. In 1997, the W3C defined the first Resource Description Framework specification. It has been used as a general method for conceptual description or modeling of information implemented in web resources by using various grammar symbols and data serialization formats. In the meanwhile, it has also been used in knowledge management applications. RDF provided a simple but powerful triple-based representation language for Universal

Resource Identifiers (URIs). For example, we want to describe a resource with the statement "there is a Person identified by "http://www.w3.org/People/EM/contact#me", whose name is Eric Miller, whose email address is e. miller123(at)example (changed for security purposes), and whose title is Dr. [8] Figure 7 is an example RDF graph from the W3C RDF Primer (www.w3.org/TR/rdf-primer), showing a representation for a person named Eric Miller.

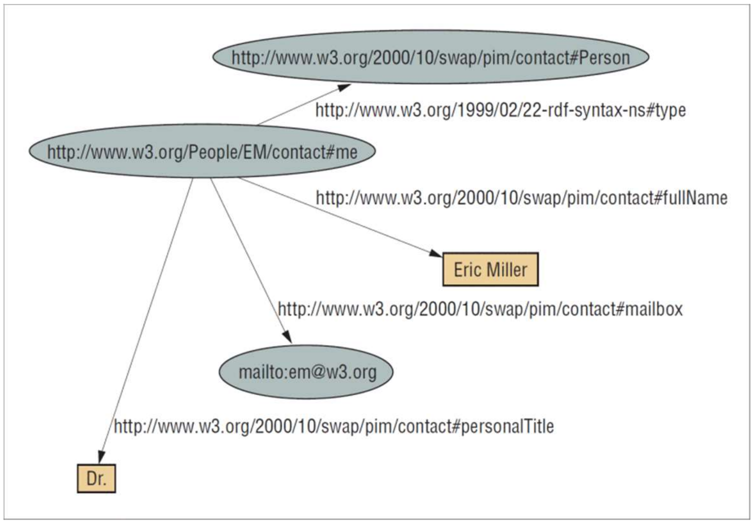
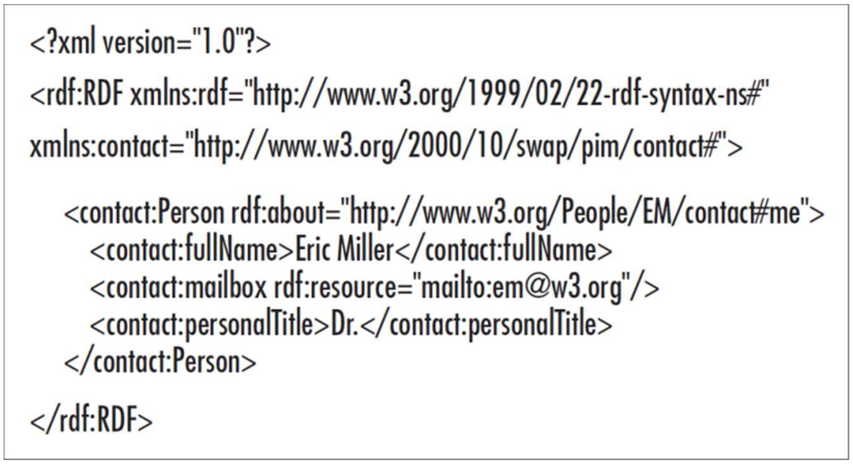

http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

Figure 7 An RDF graph representing Eric Miller.

We can use nodes and arcs to create RDF graphs. In these RDF graphs, we have individuals, such as Eric Miller, and it is identified by "http://www.w3.org/ People/EM/contact#me". We also have kinds of things, such as a person, it is identified by "http://www.w3.org/ 2000/10/swap/pim/contact#Person". In the meanwhile, we have some properties of those things, such as mailbox, can be identified by "https:// www.w3.org/2000/10/swap/pim/ contact#mailbox" and values of those properties, such as "mailto:em@w3.org" as the value of the mailbox property. RDF provides an XML-based syntax called RDF/XML for recording and exchanging graphs. As a result, the RDF can be represented in the format just like the figure 8 shows.



```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">

   <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
      <contact:fullName>Eric Miller</contact:fullName>
      <contact:mailbox rdf:resource="mailto:em@w3.org"/>
      <contact:personalTitle>Dr.</contact:personalTitle>
   </contact:Person>

</rdf:RDF>
```

Figure 8 A chunk of RDF in RDF/XML

In 2004, the RDFS (RDF Schema) was published. RDFS uses the basic RDF specification and extends it to support the expression of structured vocabularies. It provides a minimal ontology representation language, and RDFS has been widely used in research. In the same year, the Web Ontology Language (OWL) was published, which is used for those who required greater expressivity in their objects and relation descriptions. Ontologies are a formal way to describe taxonomies and classification networks, essentially defining the structure of knowledge for various domains: the nouns representing classes of objects and the verbs representing relations between the objects. [8]

To visually show the difference between RDF, RDFS, and OWL, figure 9 shows the 7-layer structures for the semantic web proposed by Tim Berners-Lee. [3]

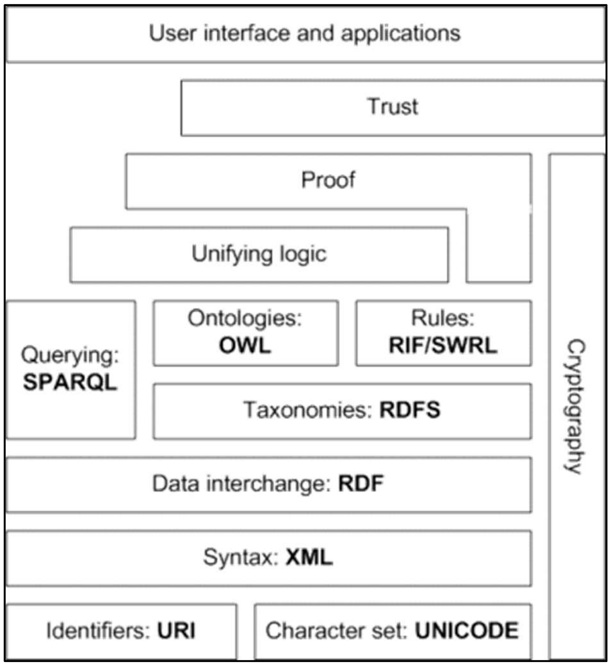


Figure 9 Hierarchical structure of semantic web

The URI / IRI in the first layer describes the resources of the web. The second layer is RDF / XML, which marks the resources on the Internet and their relationships. The XML schema describes the content and structure of the data. The third layer is RDFS / OWL, where the RDFS schema provides a specific dictionary for RDF that can be used to describe the taxonomy of classes, attributes, and their scopes. Moreover, OWL provides the semantics of domain knowledge. The fourth layer of logic is used to calculate reasoning based on RDF, RDFS, and OWL. The fifth level of certification is based on a logical description of the assessment and certification. The sixth layer of trust provides a secret relationship between users. The final layer user interface and application represent a specific application scope and different communication interfaces.

In summary, the semantic web is an intelligent network that not only understands human language but also makes communication between people and computers as easy as communication between humans. Every computer connected on the semantic web can not only understand words and concepts but also understand the logical relationship between them. They can do the work that human does. The semantic search engine is the search engine on the semantic web. It is different from a search engine with semantic analysis ability. The semantic search engine is the most direct application of semantic technology. It recognizes and processes the users' retrieval requests from the semantic level.

There is an important concept we need to introduce here to describe how a semantic search engine works, which is metadata. We can understand that metadata is structured data that machines can read it and understand it. Furthermore, we can refer to it as "the data of data." More specifically, metadata is a systematic method for describing resources and thereby improving their access. By using metadata schema, we can ensure that the machines can automatically process the metadata. So far, by using metadata, we can add the "semantic" into the search engine, then enable search engines to gain the ability to retrieve information from a semantic level. A search engine will be modified in the environment of the semantic web, and it will be annotated on the resource objects in the network; in general, a search engine should be embedded the metadata in each document or web pages. The search engine can process the users' query expression semantically, as the natural language has a logical semantic relationship. Then the semantic search engine can execute the extensive and effective semantic reasoning in the network environment to realize users' retrieval more accurately and comprehensively.

1.5 Research Challenges

As the introduction above, regardless of the methods change, the common key goal is to provide more accurate search results for their users. A search engine acts as an information intermediary when a user is facing a large amount of information. Generally, most search engines respond to a query with a 10-ranked list for each page, and the ranking reflects the search engines' estimated relevance of web pages to the query. Typically, plenty of pages will be returned by the search engine. Moreover, the users need to evaluate the results which are responded to query by the search engine. Moreover, how to efficiently decide which suggested pages should visit?

Trust is one mechanism humans use to reduce the complexity of decision making in uncertain situations [9] and maybe viewed as a fast and frugal heuristic that exploits the regularity of the information environment. [10][11] A study conducted by Helene Hembrooke, Bing Pan, and Thorsten Joachims (2007) indicated that the users only rarely view more than the first results page. [12] Most users are not aware of how the search engines "find" what they are looking for. If users trust the search engine, they will click on abstracts of higher rank. To some extent, they believed that a higher rank means more relevant results than the lower rank. Whereas, if a search engine' s ranking algorithm cannot return an accuracy result by the web pages' relevant, it will significantly affect the users' search efficiency. Therefore, the accuracy of the results ranking that a search engine responded to a query reflects the performance of the search engine.

The goal of this dissertation is to find potential methods for search engines to understand users' queries more semantically and efficiently. In this dissertation, a multi-promoting approach has been designed to improve the current traditional keyword-based search and emulate the effects of semantic search. Because annotate every web page leads to inefficiency for the search engine. Alternatively, this dissertation decided to describe a domain knowledge by using an ontology or knowledge graph. Thus, the search engine can understand users more semantically when it obtains knowledge. Whether using ontology or knowledge graph to describe a domain of knowledge, they are the symbolic and logical system, and it is difficult for the applications often involve numerical computing in continuous spaces. Thus, the main challenge of this dissertation needs to be addressed on how to convert knowledge into a machine-readable quantification result.

The knowledge is expressed very adequately by ontology. An ontology can be considered as tree-structured data, so we can calculate the semantic similarity between nodes (concepts) and get the degree of association between the keywords and web pages of the query. For the knowledge graph, we need another method to calculate the semantic similarity between nodes, because the knowledge graph is a graph structure, so we cannot process it as a tree-structure just like the ontology. Thus, the knowledge graph embedding is introduced. By embedding the knowledge graph, we can map all the nodes into space according to the relationship between them. Then we can calculate all the nodes as vectors. By using these two kinds of methods, we can convert knowledge into a machine-readable quantification result.

The core part of the remedy is how to learn the relationship between concepts from the domain knowledge efficiently. If we just rely on the semantic annotations used in the traditional semantic search. It is hard to implement because the semantic annotation is inefficiency. The models I proposed in this dissertation can optimize the process of extracting knowledge from the domain knowledge. Moreover, calculating the degree of association of concepts in this domain knowledge will help the search engine to obtain the semantic association between the query and the candidate web pages. Therefore, search engines will be able to capture the users' intention and the meaning of the web pages semantically so that users can get a better search experience.

1.6 Contributions

This dissertation discarded the semantic annotations to express the semantic similarity because markup every webpage leads to inefficiency. The remedies in this dissertation aim to improve the current traditional keyword-based search and emulate the effects of semantic search. More specifically, the proposed methods calculated the semantic similarity among the concepts in domain knowledge and added the similarity as a weight in the keyword-based search. Namely, when search engines obtain the semantic similarity among every concept in domain knowledge, it would be easy to know the degree of relevance between every pair of concepts. At the same time, once search engines have acquired domain knowledge, this knowledge can be reused efficiently through the methods provided in this dissertation. Then, when the users try to search a keyword, which is one of the concepts in the domain knowledge, the search engine will return the query results not only relate to the query keyword but also its relevant concepts in the domain knowledge. By using this approach, it can improve the search engine the capabilities to capture the conceptualizations involved in users' intention and web pages' content meaning. Because of the semantic relevance of the keywords and another vocabulary in the web pages is considered. The approaches are proposed in this dissertation not only can be used in the search engine for the web pages but also used in the information retrieval system for the documents.

1.7 The list of sub-tasks

To alleviate the limitations of the traditional keyword-based search engines, we expect the search engines could understand users' searching by content meanings rather than literal strings. This dissertation proposed ontology-based results ranking approach and a knowledge graph ranking approach to improving the current keyword-based search engine progressively.

- Topic 1: An Ontology-based re-ranking approach. This approach will rank the web pages based on the relevance between the keyword and the web pages by introducing ontology. This approach not only considers the semantic similarity in the ontology but also considers the structural factors of the concept in the ontology to improve the users' search experience on the ontology level.
- Topic 2: A Knowledge Graph-based ranking approach. As we know, the ontology has limited ability to represent the custom relation. The purpose of introducing the knowledge graph is that when the knowledge graph represents the domain knowledge, the system can still understand the semantic similarity among the

concepts and re-scoring and re-ranking the documents based on the semantic similarity. Moreover, this approach introduced the knowledge graph embedding to calculate the semantic similarity among the nodes in the knowledge graph.

1.8 Road Map

The rest of this dissertation is organized as follows. The related work of this research will be introduced in chapter 2. The methodology of the ontology-based re-ranking approach and knowledge graph-based re-ranking approach will be presented in chapter 3. In chapter 4, the experiment and its result will be introduced. The last chapter will conclude the dissertation (chapter 5).

Chapter 2 Related Work

2.1 Knowledge Representation

Knowledge representation has a long research history. [13] Moreover, the basic assumption underlying KR (and much of AI) is that thinking can be usefully understood as mechanical operations over symbolic representations. [14] Knowledge representation and reasoning have long been considered as the core issues of artificial intelligence. In general, this problem can be understood as the symbolic encoding of human knowledge and reasoning in such a way that the encoded knowledge can be read and processed by a computer to obtain intelligent behavior. On this issue, human knowledge can be varied. It can be a single person's knowledge, expert knowledge in a particular field, shared knowledge of ordinary people (common sense knowledge), or common knowledge accumulated in generations (for example, in the field of science).

In order to gather the symbols from specific domain knowledge, the system should build a computational model to access the symbols and process them. Moreover, the computational model is a knowledge base system. A knowledgebase can gather the symbolic knowledge representation in a specific domain, and the knowledge representation formalism can express the problem. For example, assume we want to represent a commonsense knowledge about Mike and Mary's photo (see figure 10).

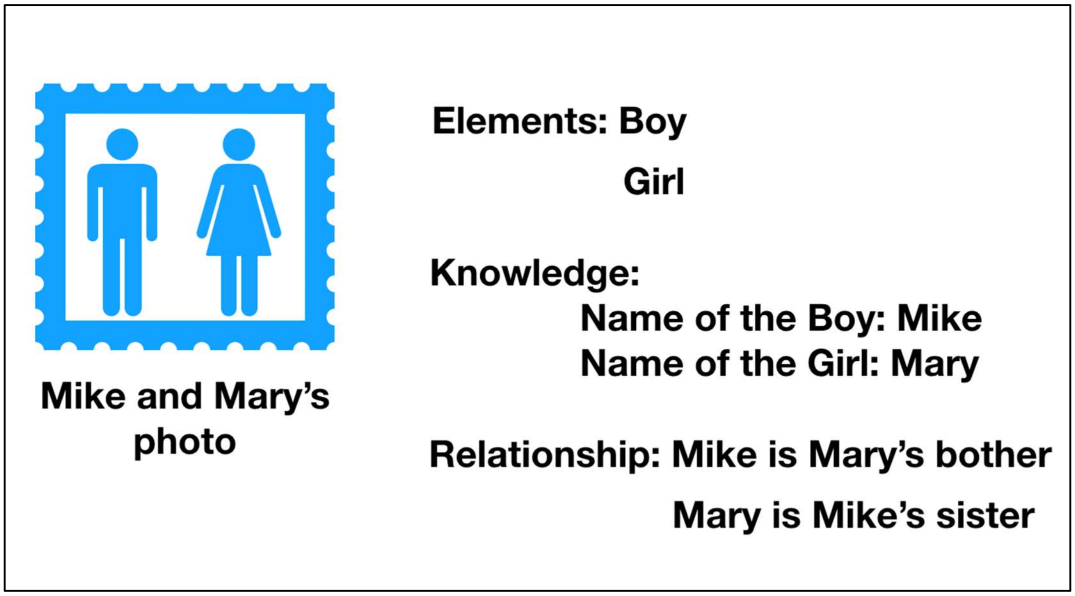


Figure 10 An example of knowledge representation

We can represent two elements in this photo, which are a boy and a girl. The knowledge about them is the name of the boy is Mike, and the name of the girl is Mary. The relationship of them is that Mike is bothered by Mary, and Mary is the sister of Mike.

Knowledge representation can express problems to be solved concerning the facts and general knowledge represented, and some requirements of the knowledge representation are listed as following: [15]

- To have a denotational formal semantic
- To be logically founded
- To allow for a structured representation of knowledge
- To have an excellent computational property
- To allow users to have maximal understanding and control over each step of the knowledge base building process and use

In a nutshell, we can understand that a knowledge representation refers to the use of computer symbols to represent knowledge in the human brain, and the process of reasoning that simulates the human brain through operations between symbols. It is the field of Artificial Intelligence dedicated to representing information about the world in a form that a computer system can utilize to solve complex tasks. The ontology and knowledge graph are examples of knowledge representation formalisms. Moreover, both of them have excellent performance when they engineer specific domain knowledge. An ontology captures the entities, their types, properties, and interrelationships between entities. A knowledge graph is a collection of entities where the types and properties have values declared for them, and where the relationships between them are mapped.

When we use the ontology or knowledge graph to represent domain knowledge, the traditional semantic search utilizes the semantic annotations to build the semantic search. More specifically, some extra data or information will be added to the webpages to describe some specific characteristics of the page. Through this approach, we can understand that the approach needs to add some data to explicitly indicate that the semantics of some words contained in this page is defined in the domain knowledge. In general, the semantic annotations are used to express semantic similarity.

2.2 Ontology

In philosophy, ontology is the study of the things that exist. In Artificial Intelligence, ontology has become a prevalent research topic since the beginning of the nineties. In general, the term ontology in Artificial Intelligence means the relationship between things on the conceptualization. For example, translating the terms in an ontology from one language to another will not change the ontology conceptually.

Ontology originates from a philosophical concept used to describe the nature of things. Gruber from the Knowledge Systems Laboratory at Stanford University first gave an ontology definition that was widely accepted in the field of information science: "Ontology is a clear specification of a conceptual model." [7] One of the reasons why ontology is essential is that its consensus on the concept of a particular field is conducive to the expression and dissemination of knowledge. In general, an ontology consists of concepts, relations, functions, axioms, and instances of five basic modeling primitives. [16] And we can use the ontology to describe and represent an area of knowledge.

One aspect of the ontology's definition needs to be clarified. The knowledge that an ontology describes or represent is just a specific area. In other words, an ontology represents a domain knowledge, and it is not to describe all knowledge. Knowledge is familiarity or understanding of someone or something, and it can refer to a theoretical or practical understanding of a subject. A classic definition of knowledge from Plato states that the knowledge must satisfy three conditions: it must be verified, correct, and believed. Therefore, we can obtain the reflection of the attributes and connections of objective things form the knowledge. Domain knowledge is knowledge of a specific discipline or field. It has all the characteristics of the knowledge, and it has a better performance for the expression of a particular area. As each industry has its unique domain knowledge, therefore, domain knowledge is valid knowledge, and it is useful for representing an area of a specialized field.

From the definition mentioned above, another important concept in the Semantic Web is taxonomy. The Semantic Web is an extension of the World Wide Web through standards by the World Wide Web Consortium (W3C). It can provide the machine-readable definitions of all kinds of things and their relationships among them. The ultimate goal of the Semantic Web is to enable machines that can better understand the human intention. In order to do that, the Semantic Web must achieve the information from the linked data. Moreover, taxonomy and ontology are often used to organize the linked data, and they can be used interchangeably. However, they are different concepts. Taxonomy is a set of definitions that has a hierarchy structure. It starts from the most general description of one thing or concept and details it as the term goes down. For example, as shown in figure 11, the Audi A4 is a specific description of the vehicle. And an ontology describes a concept not only by its position in a hierarchy but also its relationship to other concepts. For example, the Audi A4 would also be associated with the concept of the brand Audi or the Volkswagen Group. Both the taxonomy and the ontology have the tree structure if we refer to taxonomy as one "tree," and an ontology should be the "forest."

According to the characteristics of the ontology, the ontology becomes a powerful tool to represent the domain knowledge. Because ontology can provide a common and shared definition in a domain, it can make the knowledge representation more efficient, as it provides a way to reuse the domain knowledge. The ontology also provides a machine-readable definition, as it has a way to encode the knowledge, such as RDF, OWL. With these advantages of the ontology as evidence, it could be one of the most critical approaches to improve a search engine to understand users' searching by content meanings.

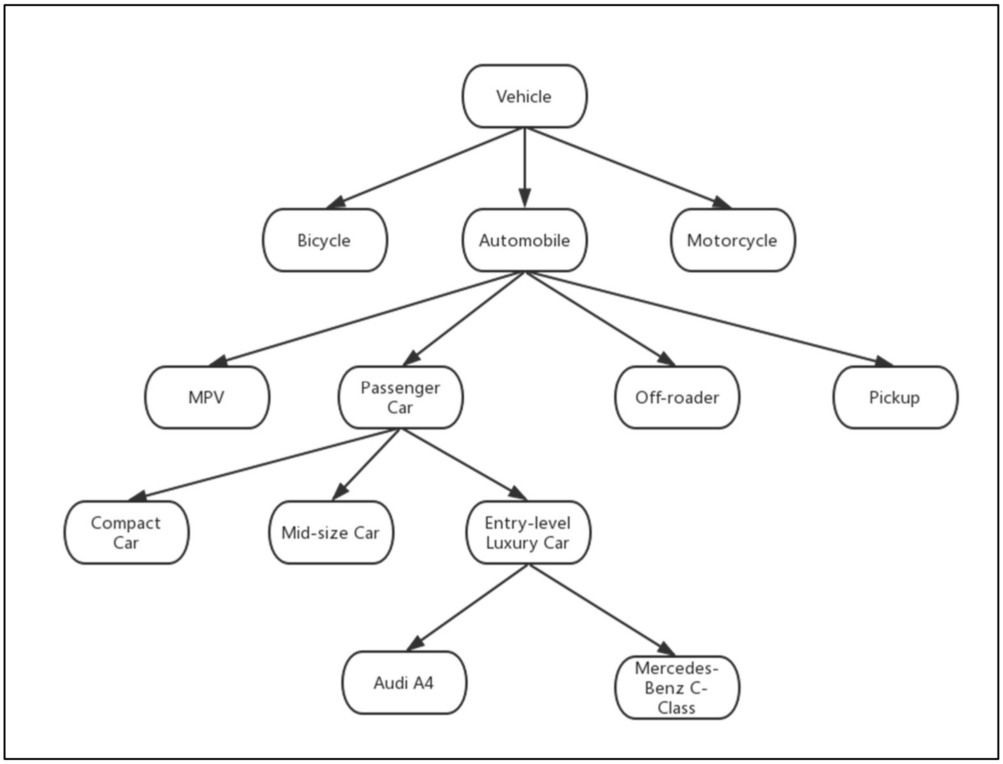


Figure 11 The illustration of Taxonomy

The application of ontology mainly involves two aspects: first, as a tool, the ontology can provide knowledge sharing and reuse in the knowledge layer; second, the ontology can be applied in the information system. In the information system, the ontology can be implemented in information organization, information retrieval, and complex information system interoperability issues.

2.3 Protégé

Protégé (https://protege.stanford.edu/) is a free, open-source ontology editor and a knowledge management system. The Protégé is mainly used for the construction of ontology in the semantic web, and it is the core development tool for ontology construction in the semantic web. Protégé provides a graphic user interface to define ontologies. It also provides the construction of ontology concept classes, relationships, attributes and instances, and shields the concrete ontology description language. Users only need to construct the domain ontology model at the conceptual level. The software has a self-configurable data input mode, which can convert the internal representation of protégé into various forms of text representation, such as XML, RDF (S), OIL, DAML, DAML + OIL, OWL and other system languages. Figure 12 is an example of the Protégé users' interface.

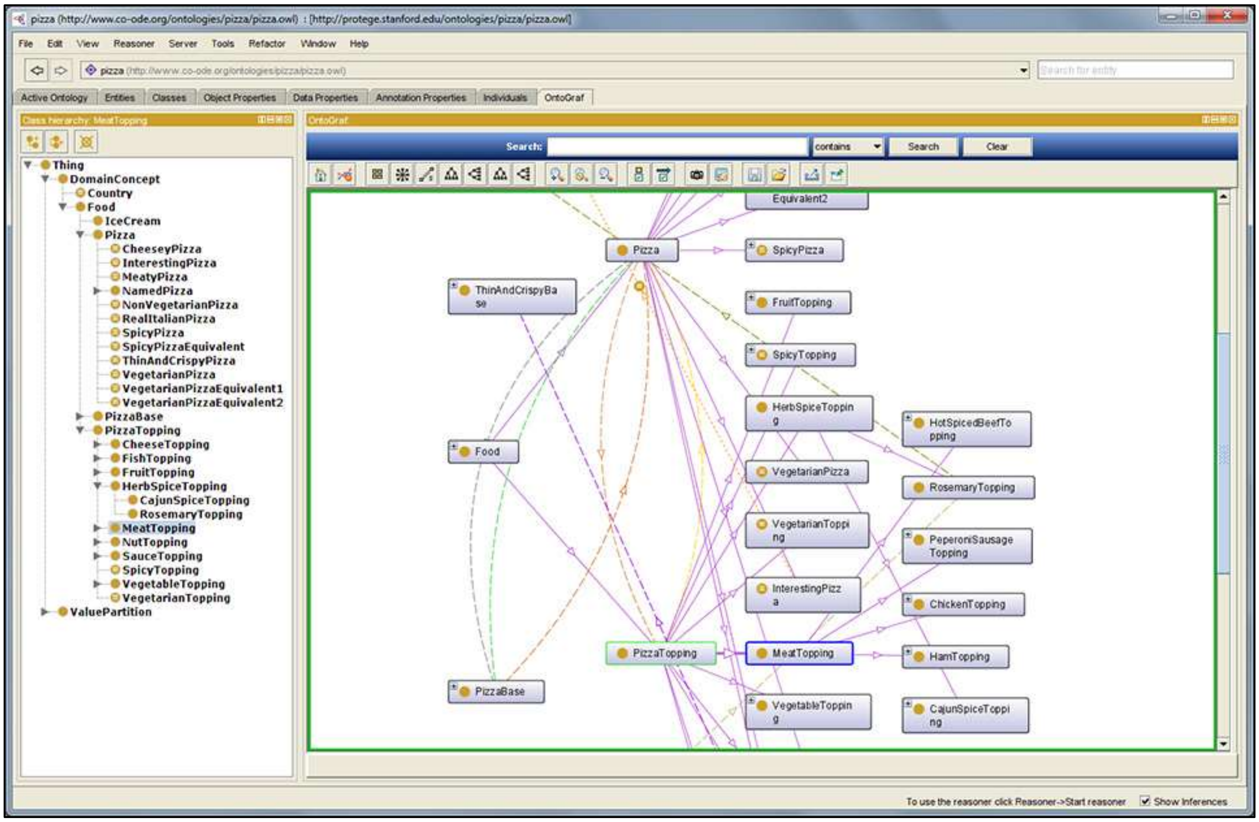


Figure 12 Demo of the Protege user interface

2.4 Apache Jena

Apache Jena is an open-source Semantic Web framework for Java. It can be used for application development in the Semantic Web. The Apache Jena framework mainly includes:

1) Read and write RDF in RDF/XML triples. The Resource Description Framework (RDF) is a standard for describing resources (technically the W3C recommendation), and Apache Jena can create, read, write, and query RDF models.
2) Like the ontology operations such as RDFS, OWL, DAML+OIL, the Jena framework includes an Ontology Subsystem that provides an API that allows the processing of RDF-based ontology data. Namely, it supports OWL, DAML+OIL, and RDFS. The Ontology API combines with the reasoning subsystem to extract information from a specific ontology. Apache Jena also provides a Document Manager (OntDocumentManager) to support document management for imported ontology.
3) Use the database to save data. Apache Jena allows data to be stored on the hard drive, either in an OWL file or in a relational database.
4) Query model. Apache Jena provides the ARQ query engine, which implements the SPARQL query language and RDQL so that Apache Jena can query different models. Besides, the query engine is associated with a relational database, which enables higher efficiency when the query is stored in an ontology in a relational database.
5) Rule-based reasoning. Apache Jena supports simple rule-based reasoning. Its reasoning mechanism is to import inference reasoners into Apache Jena. When creating a model, the reasoner is associated with the model to achieve reasoning.

2.5 Semantic Similarity

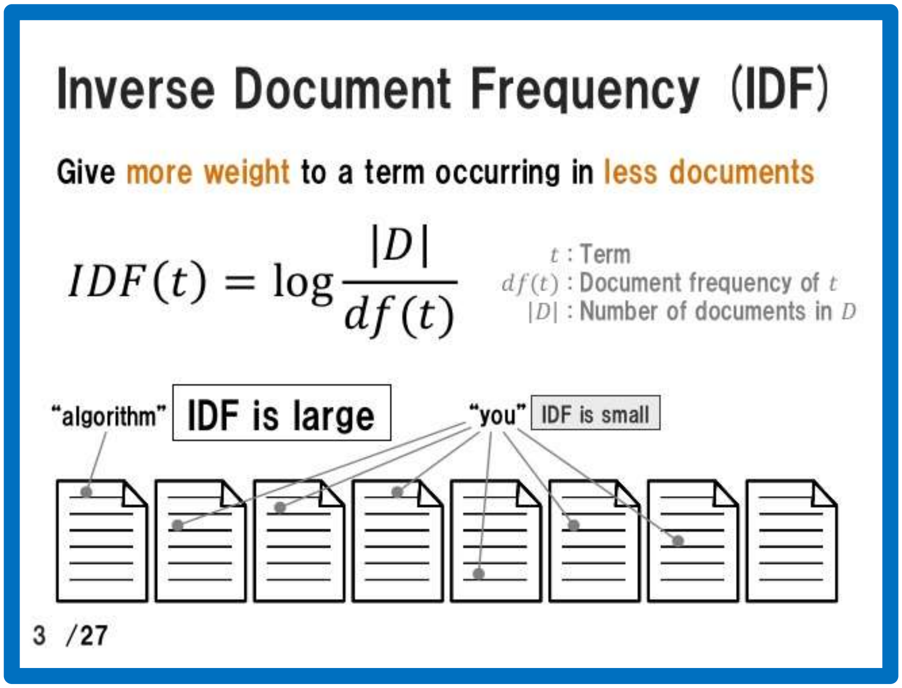
Semantic Similarity is a metric defined over a set of documents or terms; it is a measure of the degree of closeness between two things. For example, cars and gasoline seem to be more closely linked than cars and bicycles. Namely, the distance between the terms is based on the relevance of their meaning or semantic content. Semantic Similarity is widely used to compare the semantic entities such as units of language, concepts, or even semantically characterized instances. [17] Semantic Similarity is widely used in many practical applications, particularly in natural language processing (NLP), including semantic information retrieval, keyword extraction, and document summarization, where it can be used to quantify the relations between words or between words and documents. [18] Information retrieval techniques have an interest in semantic relatedness measures as their incorporation in the retrieval process allows the identification of meaningfully related but lexically unique content. [19]

One of the earliest applications of text similarity may be the vector model in information retrieval, where the document most relevant to the input query is determined by sorting the documents in the collection in the reverse order of Similarity to the given query. [20] In this dissertation, the semantic similarity is used to measure the relevance between the query and the other concepts. Then the search engine could understand which concept is the most similar to the query. Accordingly, the search engines can obtain the relevant web pages and rank them by both the keyword’s frequency and the semantic meaning in the domain knowledge. By using this approach, we can utilize the semantic search to mine the users’ internal intention without semantic annotation. Computationally, semantic Similarity can be estimated by defining a topological similarity.

2.6 TF-IDF

TF-IDF, Term Frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. [21] The TF-IDF is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. Because the TF-IDF is an excellent approach to measure the degree of correlation between a file and a user query, the high word frequency within a particular file, and the low file frequency of the word in the entire file set, can produce a high weight TF-IDF. Therefore, TF-IDF tends to filter out common words and retain important words and can help to adjust to the fact that some words appear more frequently in general. Today, TF-IDF is one of the most popular term-weighting schemes; 83% of text-based recommender- systems in digital libraries use TF-IDF. [22]

Using IDF to measure how important a word in a document is easy to understand the term frequency. The main idea is to give more weight to a term occurring in fewer documents. For example (see figure 13), the word “you” appears in many documents, and the word “algorithm” appears in one document. Therefore, the word ‘algorithm’ has a better different ability than the word “you,” then the “algorithm” is more important than “you.” By using this approach, it is easy to filter out the terms that have a high frequency in the documents but have low-quality, such as “you,” “and.”



Inverse Document Frequency (IDF)

Give more weight to a term occurring in less documents

$$IDF(t) = \log \frac{|D|}{df(t)}$$

t : Term
$df(t)$: Document frequency of t
$|D|$: Number of documents in D

"algorithm" IDF is large

"you" IDF is small

3 /27

Figure 13 A demonstration of IDF

In a given document, the term frequency (TF) refers to the number of times a given the word appears in the file. This number is usually normalized (generally, the numerator is less than the denominator is distinguished from the IDF) to prevent it from being biased towards large files. (The same word may have a higher word frequency in a log file than a small file, regardless of whether the word is essential or not.) For the word t_i in a particular file, its importance can be expressed as:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

In Equation 3, $n_{i,j}$ is the number of occurrences of the word t_i in the file d_j, and the denominator is the sum of the occurrences of all words in the file d_j. The inverse document frequency (IDF) is a measure of the universal importance of a word. The IDF of a particular word can be obtained by dividing the total number of files by the number of files containing the word and then taking the logarithm of the resulting quotient.

$$IDF_i = lg\frac{N}{dfi} \quad (4)$$

N is the amount of the documents set, dfi is the number that the word t_i appears at least once in the document. And the TF-IDF (term frequency-inverse document frequency) can be represented as:

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \quad (5)$$

The main idea of the TF-IDF (term frequency-inverse document frequency) is if a word or phrase appears in an article with a high frequency (high TF value). It is rarely found in other articles (high IDF value), the word or phrase can be considered as having the excellent ability of the class distinguishing and is suitable for classification. For example, if we assume four words appear in three documents, and the value of each word is listed in the following table.

Table 1 The example in the TF-IDF calculation

	Doc 1	Doc 2	Doc 3
Car	27	4	24
Auto	3	33	0
Insurance	0	33	29

Best	14	0	17

For the word “Car” in these three documents, we got

$$TF_{car,Doc1} = \frac{27}{27+3+0+14} \approx 0.61,$$

$$TF_{car,Doc2} = \frac{4}{4+33+3} \approx 0.06,$$

$$TF_{car,Doc3} = \frac{24}{24+0+29+17} \approx 0.34,$$

$$IDF_{car} = lg\frac{3}{3} = 0.$$

Therefore,

$$TFIDF_{car,Doc1} = 0.61 \times 0 = 0,$$

$$TFIDF_{car,Doc2} = 0.06 \times 0 = 0,$$

$$TFIDF_{car,Doc3} = 0.34 \times 0 = 0.$$

And for the word “Auto” in these three documents, we got

$$TF_{auto,Doc1} = \frac{3}{27+3+0+14} \approx 0.07,$$

$$TF_{auto,Doc2} = \frac{33}{4+33+33+0} \approx 0.47,$$

$$TF_{auto,Doc3} = \frac{0}{24+0+29+17} = 0,$$

$$IDF_{auto} = lg\frac{3}{2} = 0.18.$$

Therefore,

$$TFIDF_{auto,Doc1} = 0.07 \times 0.18 = 0.0126,$$

$$TFIDF_{auto,Doc2} = 0.47 \times 0 = 0.0846,$$

$$TFIDF_{car,Doc} \ \ = 0 \times 0.18 = 0.$$

From the value of the TF-IDF, we can understand that the word “auto” has a higher weight than the word “car.” (the word “auto” is more important than the word “car” in the file set).

2.7 Knowledge Graph

The development of the knowledge graph is based on the maturity of the technologies such as knowledge mapping, citation analysis, information resources, information science, and quantitative analysis. The researches and applications of the knowledge graph have created a boom in academics and industrial because of technological improvement.

The essence of the knowledge graph is to represent knowledge better. The concept of the knowledge graph is not new, and the idea behind it can be understood as the extension of Ontology. An ontology consists of interconnected nodes and edges. The nodes represent the concepts or the objects, and the edges represent their relationship. In terms of expression, the knowledge graph and the ontology are similar. The knowledge graph uses the nodes to represent the entities and uses the edges to represent the relationships among the entities. However, the ontology focuses on describing the relationship between the concepts, while the knowledge graph focuses more on describing the association among the entities. To some extent, the knowledge graph obtains more abundant information in the knowledge representation than ontology.

From an abstract perspective, an ontology is more abstract than a knowledge graph. For example, if we want to create a knowledge base or a knowledge graph for a library's book. First, we should classify the books, and this classification can be considered as an ontology. For example, books can be divided into computer science and electronics. Computer science can be classified into networks and artificial intelligence. With this classification, we can divide the books into each category, and this level of classification can be considered as an ontology. Figure 14 is a demonstration of the book's ontology.

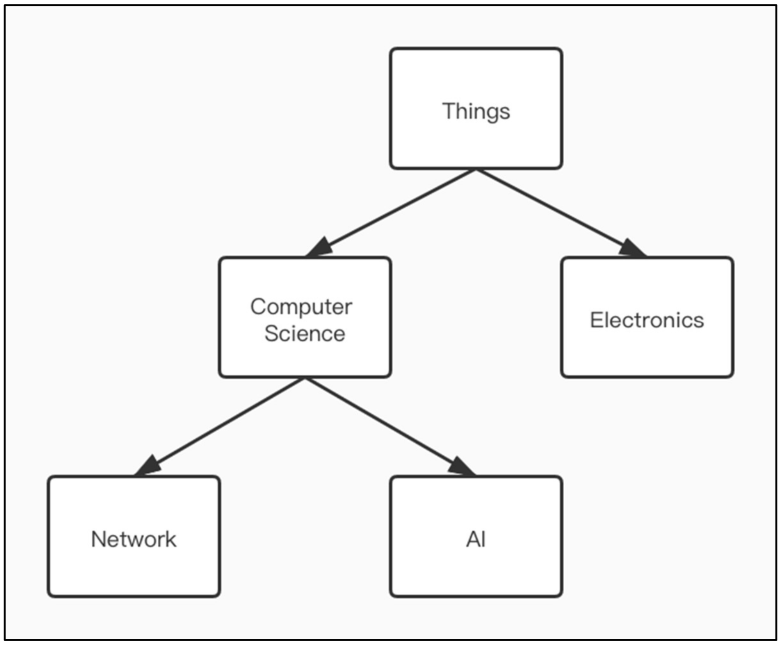


Figure 14 A demonstration of ontology

The knowledge graph can be considered as the "upgraded version" of ontology because the knowledge graph can not only express the relationships among the concepts but also can describe the relationship among the entities. Compared with the ontology, the knowledge graph has a more abundant relational expression, and we call this relationship the customer relationship. For example, “Zero to One” is a book based on Peter Thiel’s lecture notes on the video and handouts of the “Startups” course at Stanford University. In the ontology, this book can only be considered as an instance of a computer science book. However, if we use the knowledge graph to represent it, we can express many entities, as well as the relationships among the entities. Such as the author of this book, the University of the author. Figure 15 is a demonstration of a knowledge graph.

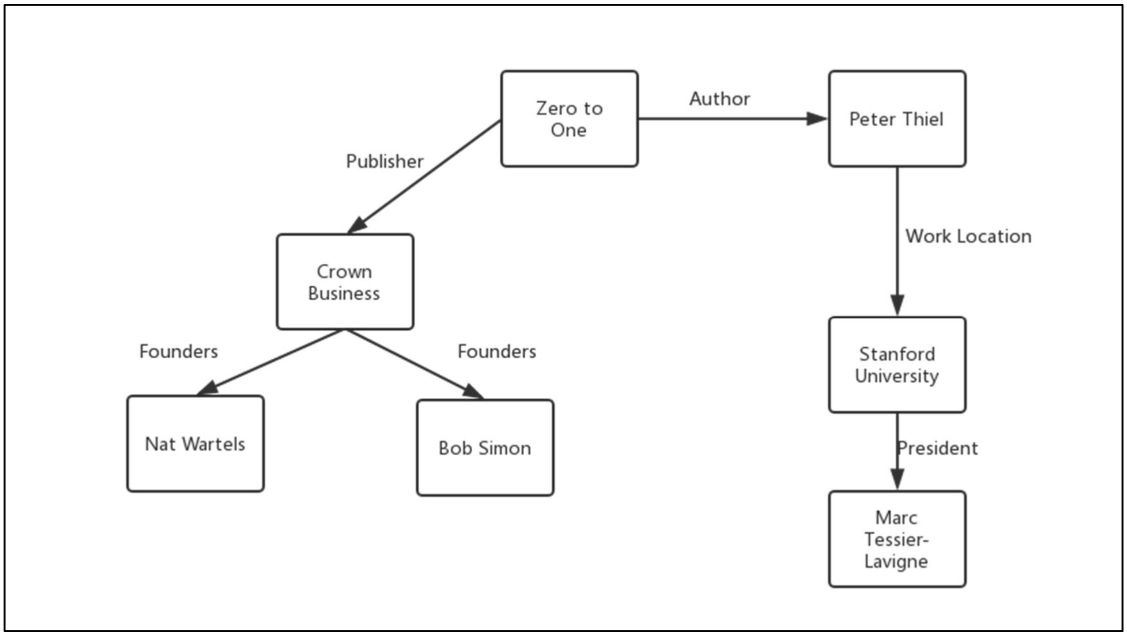


Figure 15 Demonstration of a knowledge graph

2.8 Pace Protégé

In the previous chapter, Protégé is an open-source tool developed by Stanford University Medical Informatics. It is an open-source tool and used as an ontology editor, and it provides a suite of tools to construct the domain model through various formats. Also, using plug-ins for adding further functions makes it flexible. These plug-ins, such as importing and exporting ontology language specifications like XML, RDF, RDFS, OWL, and different types of reasoners, are available. Web Ontology Language (OWL) is used by domain experts to encode knowledge.

OWL primarily only supports the subClassOf (is-a or inheritance) relation. The variety of other relations, such as “part of” relations, are essential for representing information in various fields, including all engineering disciplines. The current syntax of OWL does not support the declaration and usage of new custom relations. Representing the part-whole relationship is a widespread issue among people who want to develop ontologies for the semantic web. The “partOf” relationship is one of the basic fundamental primitives of the universe. Many applications are required to show this kind of relationship in the real world. RDF schema and OWL does not provide any built-in primitives to support the part-whole relationship. In Workarounds to emulate custom relations do exist, but they add syntax burden to knowledge modelers and do not support precise semantics for inference engines.

Pace University extended OWL with minimal syntax extension to allow domain experts to declare custom relations with various mathematical properties. [18] The resulting knowledge representation is called Knowledge Graphs. Pace University has also extended Stanford University’s project for Protégé v5 extended version of Protégé and OWLViz allows the users to declare new custom relations with distinctive attributes and apply the knowledge representation to be visualized in a knowledge graph. Then the domain experts can be visually declared custom relations and encode knowledge. Figure 5 shows the extension of the “partOf” relations and custom relations such as “include,” “ref,” “implemented,” and “implementedBy.” The entities tab allows users to view the class hierarchy for the Web Tutorial domain ontology. In this pane, users can add new classes, subclasses, and remove classes. The relation’s pane allows users to add their custom relations and remove custom relations. In Figure 16, Pace Protégé GUI interfaces with custom relations and related to feature show how to relate classes to the “include” relations for the class “WebTechnology.” The relations are used to relate these relations to each class in the domain. Once this has been completed, we can then use Pace’s extended customized version of OWLViz to visualize our knowledge graph with these new custom relations.

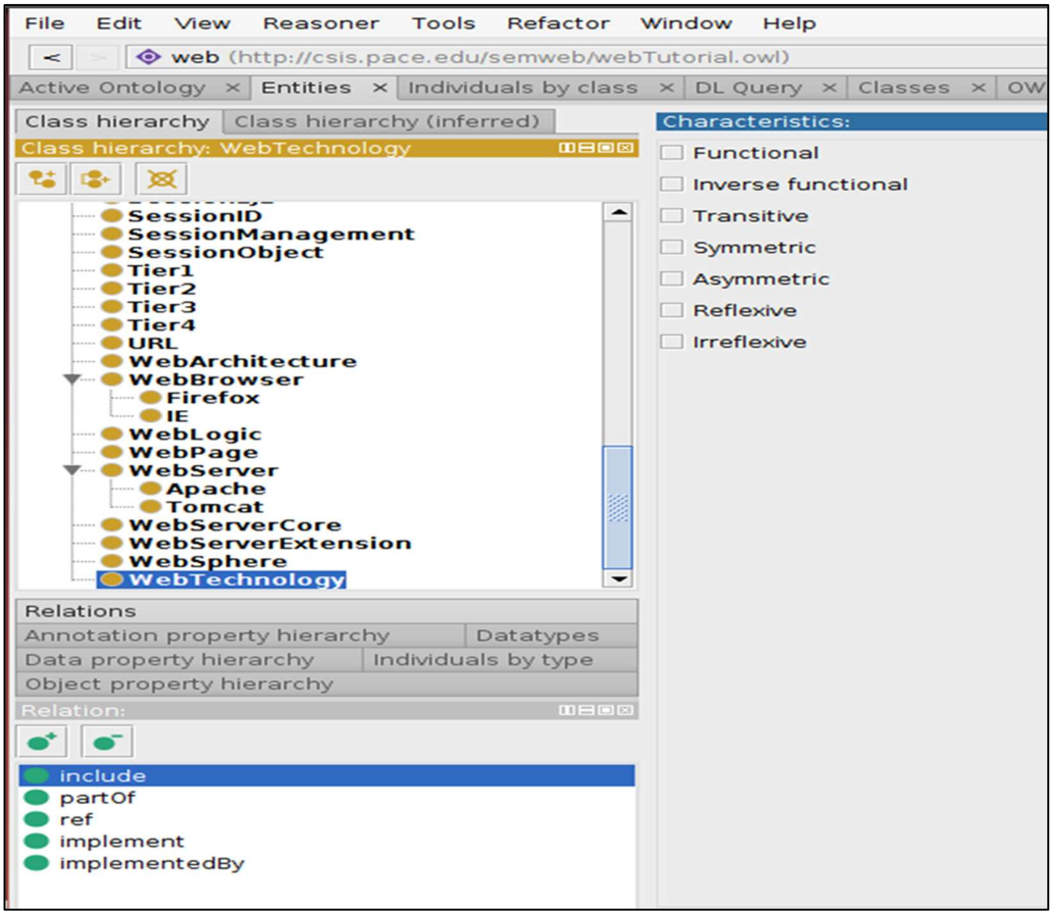


Figure 16 Pace Protege GUI Interface with Custom Relations Features

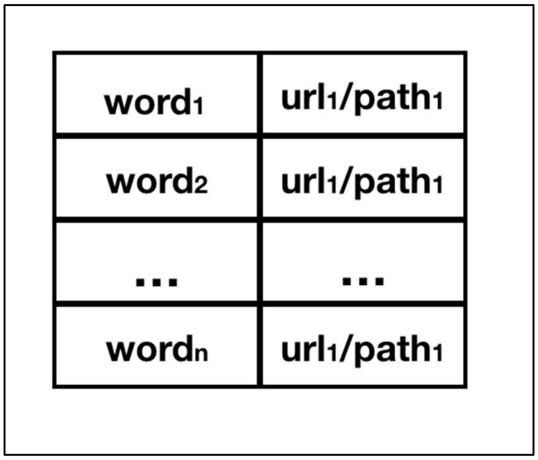
2.9 How a traditional Search Engine works

As the introduction in the previous chapter, if we want to collect the information from a vast number of documents or data, we need search engines to realize the retrieval of massive information. Next is the introduction of how traditional search engines work and retrieve user queries from vast amounts of data.

2.9.1 Index table construction

In order to be able to retrieve information more quickly for potential users, search engines often start with an extensive index table, and this procedure also being called an indexation process. Usually, the indexation process will be conducted by a crawler, which is a software that can collect the documents or webpages and construct the index table. The crawler will begin to construct the index table from a seed URL and do the following:

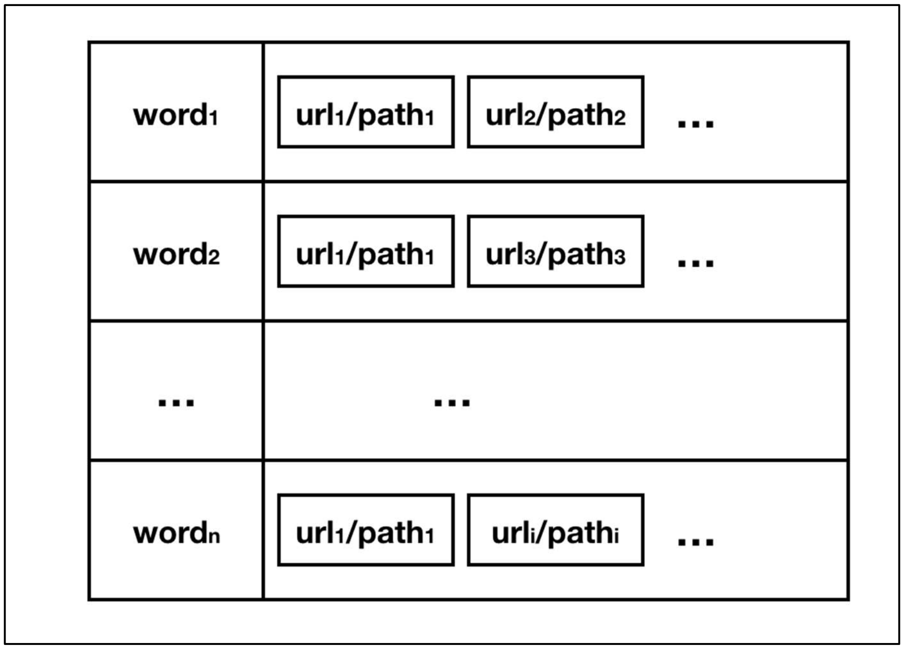
- Read every single word in the seed URL and create an index table for it. For example, if we denote that the url_1 for the seed URL, the crawler will read every single word ($word_1$, $word_2$, …, $word_n$), and assign the url_1 for these words (see the figure 17).



word$_1$	url$_1$/path$_1$
word$_2$	url$_1$/path$_1$
...	...
word$_n$	url$_1$/path$_1$

Figure 17 The initial index table

- Download another webpage which is linked by the seed webpage. In the documents database, we can think of the link in the path of the document.
- Read every word on the webpage or document and add them to the initial index table.
- When all the webpages or documents are processed, the final index table will look like the following figure.



word$_1$	url$_1$/path$_1$	url$_2$/path$_2$	...
word$_2$	url$_1$/path$_1$	url$_3$/path$_3$	...
...	...		
word$_n$	url$_1$/path$_1$	url$_i$/path$_i$	...

Figure 18 Final index table

2.9.2 The query procedure

When an index table prepared, the user can start their search on it. A search engine collects the URLs or the document's paths according to the users' query. Then, the search engine ranks these webpages or documents base on the association between the query and the collected webpages or documents. The frequency of the query calculated the association occurs on the webpage or document. (as shown in the following figure).

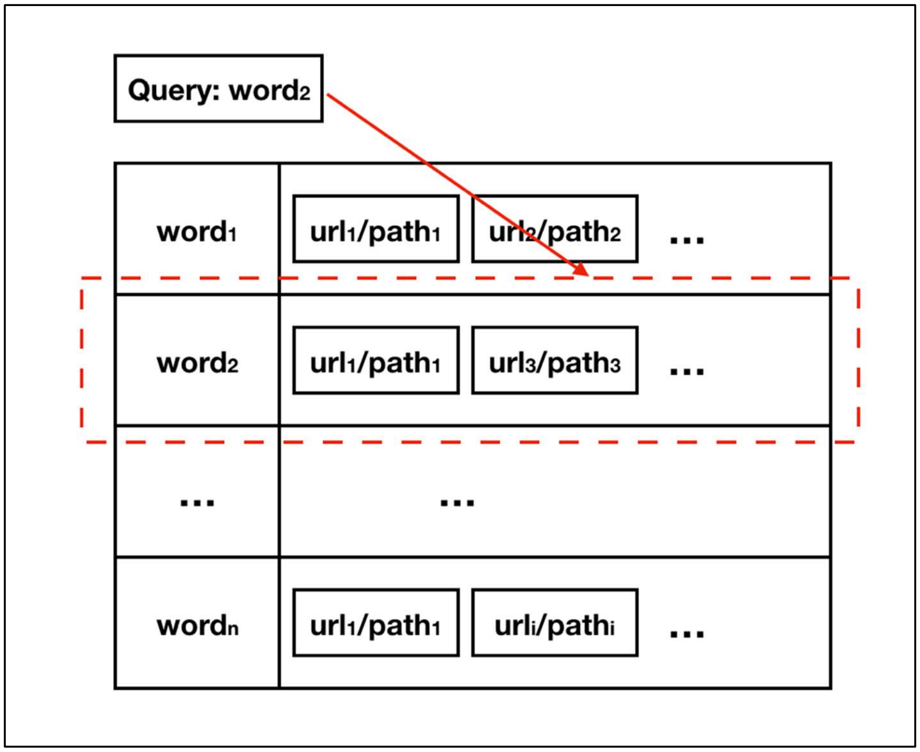


Figure 19 Search engine return the results for the query word$_2$

2.10 How a traditional semantic search engine works

Due to the low ability to explore the users' intention in traditional search engines, this dissertation introduced the semantic search engine to alleviate the limitations. For example, [23] if a user wants to search SLR (single-lens reflex) on the web, his or her real purpose is actually to get more information about single-lens reflex or photography. However, the traditional search engine may have no idea about the SLR, and it will return to the user the webpages which contain the string SLR. More specifically, suppose the word SLR has some links in the index table, such as:

- www. cheapCameras.com
- www.buyItHere.com

- Other links

Then, the query in the index table is shown in figure 20.

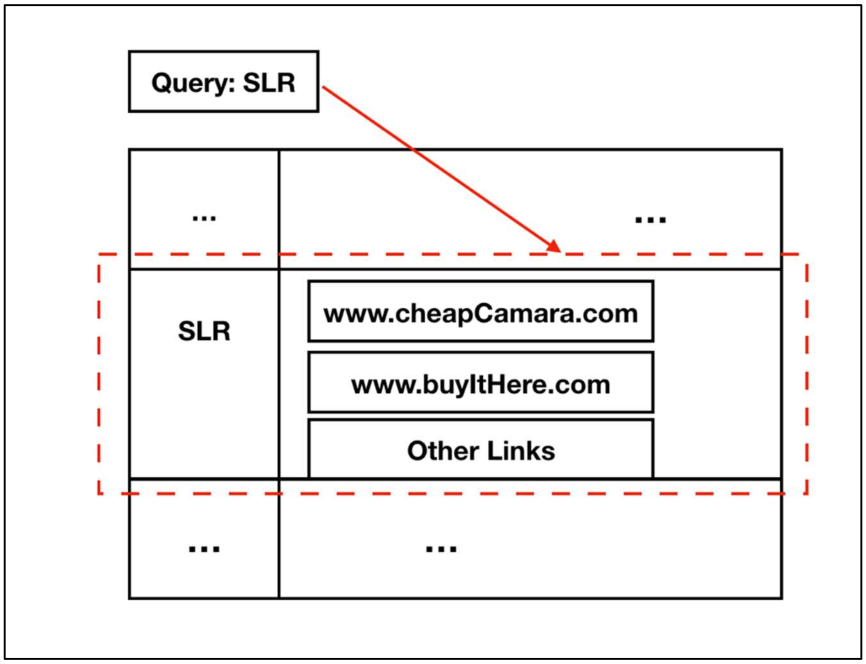


Figure 20 SLR as query in the index table

However, the returned links may all be sales sites, because the frequency of query keywords appearance is high on these sites. Furthermore, the traditional search engine is easily missing some webpages that may match the users' intentions, due to lacking the query keyword or the low frequency of the keyword (for example, the web www.goodPhoto.com). This section is the introduction of how a traditional semantic search engine work. [23]

- First, we need to add domain knowledge to the web or database. Usually, we use the knowledge base to save the domain knowledge, for example, we can create a small vocabulary for the SLR (see figure 21)

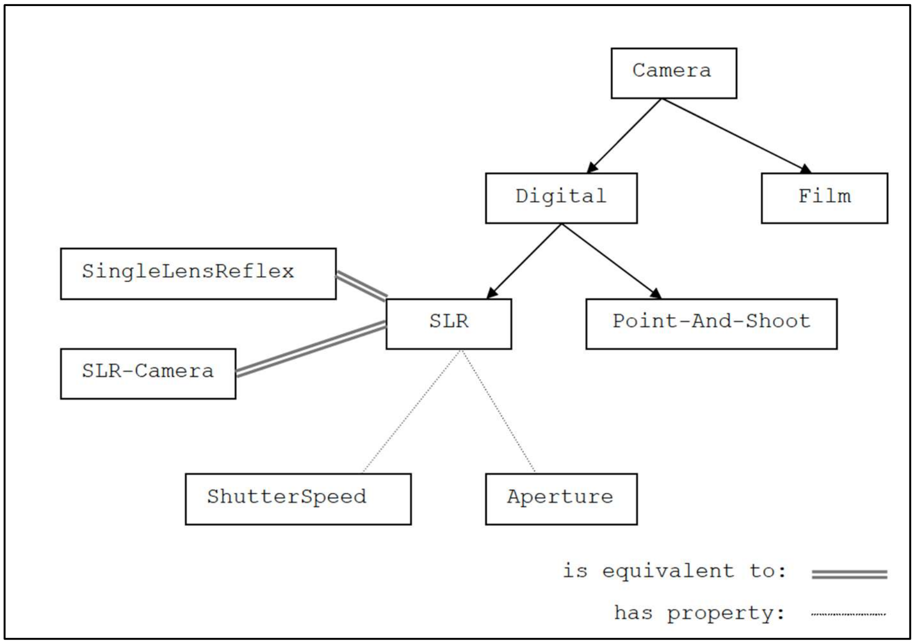
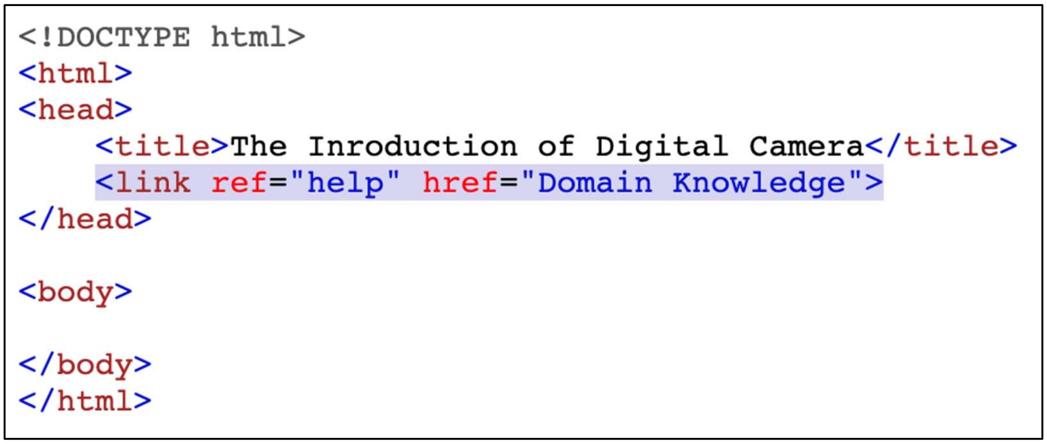


Figure 21 A domain knowledge of SLR

- Second, the search engine should markup the webpages or documents. When the semantic search engine obtained the knowledge, we need to let the search engine have the ability that identifies the webpage like www.goodPhoto.com (the example we used previously) that is highly relevant to the query SLR. Therefore, we need to annotate the webpage by using domain knowledge. For example, we can add the annotation on the webpage www.goodPhoto.com like figure 22.



```
<!DOCTYPE html>
<html>
<head>
    <title>The Inroduction of Digital Camera</title>
    <link ref="help" href="Domain Knowledge">
</head>

<body>

</body>
</html>
```

Figure 22 Add an annotation on a webpage

- Build an index table. In the semantic search engine, we will have a smarter crawler as the domain knowledge is introduced. Compared with the traditional search engine, the crawler of the semantic search engine will not only use each word appearing on the webpage or document as the basis for establishing the index table when passing through each webpage or document. By introducing domain knowledge, we have labeled the domain knowledge in every web page or document, the crawler of the semantic search engine can know the relationship between the various concepts in the domain knowledge. The crawler will add those assigned the URL or file path to the corresponding index while building the index table. In general, the difference of this step between the traditional search engine and the semantic search engine is, all the related links or paths of the documents will be added to the corresponding words (see figure 23).

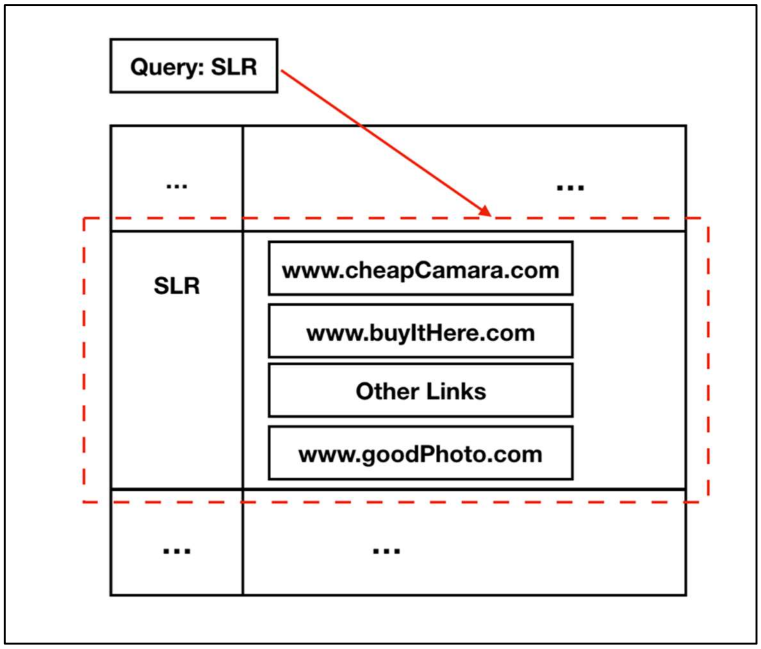


Figure 23 The index table with www.goodPhoto.com

Then, the traditional semantic search engine is ready to use. The semantic search engine will collect the URLs or the document's paths corresponding to the users' query. Then, the search engine will rank these webpages or documents based on the association between the query and the collected webpages or documents. Moreover, the association was calculated by the frequency which the query occurrence on the webpage or document. (as shown in figure 23).

2.11 Conclusion

In this chapter, some knowledge of the traditional search engine and semantic search are introduced, including the concepts related to knowledge expression and description of how knowledge is stored. Some software that deals with knowledge expression is also briefly introduced in this chapter. Moreover, some methods for calculating the degree of relevance are introduced in this chapter, such as semantic similarity and TF-IDF (term frequency-inverse document frequency). Finally, this chapter details the construction and retrieval principles of traditional search engines and semantic search engines. Based on this research, we can realize that there is still room to improve the current solution. In the next chapter, the dissertation will elaborate on the methods to solve the problems.

Chapter 3 Methodologies

In order to solve the problems encountered by traditional search engines, we need to introduce domain knowledge into search engines, so that search engines can have the ability to search semantically. The traditional semantic search uses semantic annotation to complete the users' search. Specifically, each webpage in the index table is labeled according to the knowledge from the domain knowledge. However, the method will lead to the decline of search efficiency.

3.1 Ontology-based Re-ranking approach

Ontology is a very effective way of expressing knowledge. [24] An ontology can be considered as tree-structured data so that computers can efficiently calculate the semantic similarity between nodes (concepts) and get the degree of association between the keywords and web pages of the query. Therefore, with this remedy, search engines can efficiently and accurately capture the semantic information behind the user's query and complete the semantic search. The main challenge of this dissertation needs to be addressed how to convert a knowledge (ontology and knowledge graph) into a machine-readable quantification result. The second task is to calculate the degree of association of concepts in the domain knowledge. Thereby improving the current traditional keyword-based search and emulating the effects of semantic search would be the expected result

3.1.1 The Workflow

Based on the traditional search engine, the first approach is to use the ontology to improve the search experience on the semantic level. [25] Namely, some prepared web pages in an index table have already been prepared, and the highly relevant web pages will be returned from the ontology-based ranking model. The model includes five parts:

- Create the domain knowledge base;
- Get the candidate documents or web pages from the dataset;
- Calculate the semantic similarity between the concepts in the ontology;
- Score the candidate web pages by the semantic similarity and the term's TF-IDF weight;
- Rank the candidate web pages by the score, which are generated in step 4 and return the result.

Figure 24 shows the framework of the model.

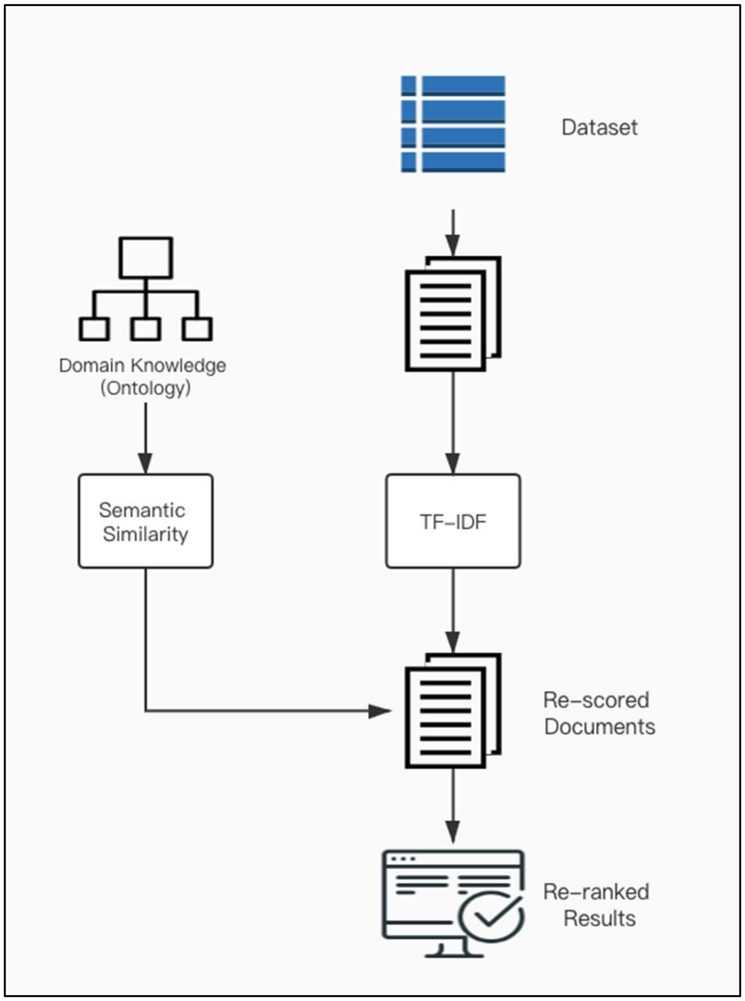


Figure 24 The Framework of the Ontology-based re-ranking approach

3.1.2 The semantic similarity calculation

The semantic similarity between the concepts in the ontology can be considered as the semantic distance, semantic coincidence, and the level difference.

a. Semantic Distance: We assume that X and Y are two nodes (or concept) in the ontology, and the shortest path between X and Y is the Semantic Distance, referred to as Dis (X, Y). Semantic distance is an essential element when we compute the Semantic Similarity. When the distance between two conceptual paths are long, the

Semantic Distance is long, and the Semantic Similarity is small. For example, based on the example illustrated in chapter 2 (Figure 11), we can calculate that the semantic distance between the Audi A4 and Benz c class is 2. Moreover, the semantic distance between Audi A4 and pickup is 4; that is, the semantic similarity between Audi A4 and Benz c class is high (both are luxury cars), and the Audi A4 and Pickup has a low semantic similarity (different types). Figure 25 shows an example of the semantic distance.

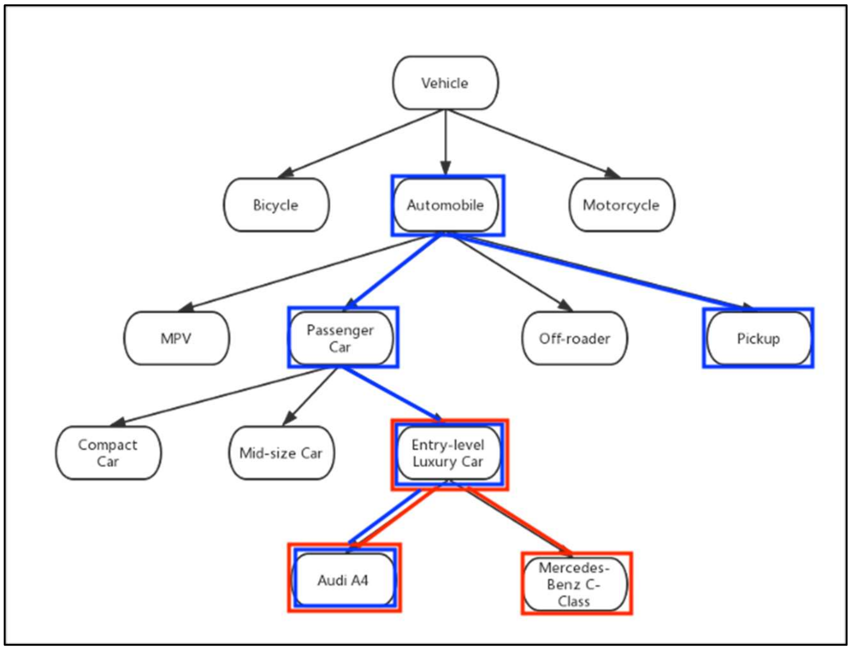


Figure 25 Example of the Semantic Distance

b. Semantic Coincidence: We can assume that X and Y are two nodes (or concept) in the ontology, N(X), and N(Y) represent that the number of nodes to reach the root node R from X and Y respectively.

$$Semantic\ Coincidence = \frac{|N(X) \cap N(Y)|}{|N(X) \cup N(Y)|} \tag{6}$$

The Semantic Coincidence represents the same degree between the two concepts. For example,

- N (Audi A4) ∩ N (Benz c class) = 4;
- N (Audi A4) ∪ N (Benz c class) = 6;

the semantic coincidence of Audi A4 and Benz c class is 4 over 6 (0.67).

- N (Audi A4) ∩ N (pickup) = 2;
- N (Audi A4) ∪ N (pickup) = 6;

the semantic coincidence of Audi A4 and pickup is 2 over 6 (0.34). We can understand that the semantic similarities between the Audi A4 and Benz c class are higher than the semantic similarities between the Audi A4 and Pickup.

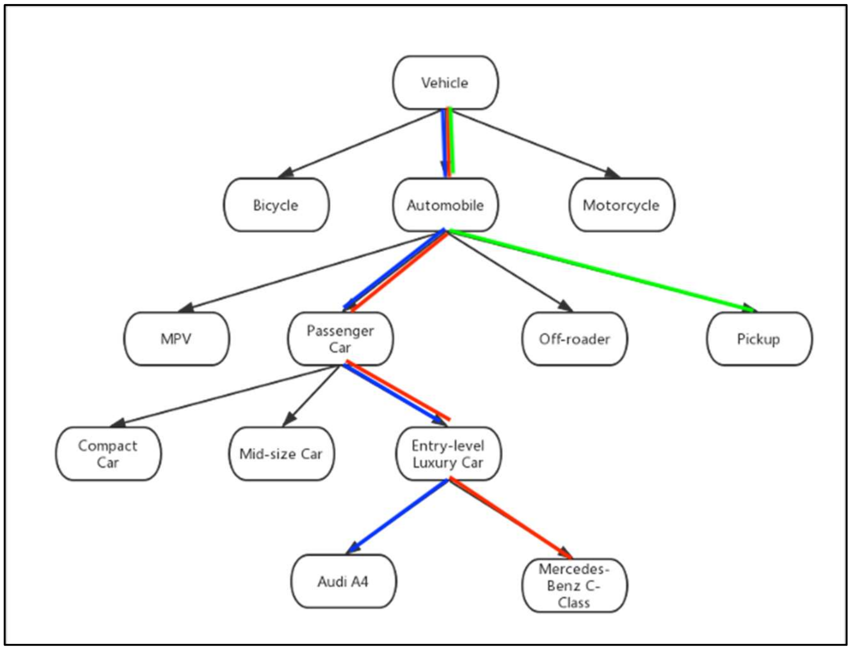


Figure 26 The example of semantic coincidence

c. Level Difference: We can assume that X and Y are two nodes (or concept) in the ontology, L(X) and L(Y) are the levels where the concepts X and Y are, the Level Difference is |L(X) – L(Y)|. The information number of different concepts is not the same if they are on a different level at the ontology tree. The bigger the Level Difference is, the lower the Semantic Similarity is. For example, Audi A4 and Benz c class are in the same level of the ontology tree, the level difference is 0, and the level difference between Audi A4 and pickup is 2. From the human understanding, Audi A4 and Benz c class are not only a kind of car but also are the instance of the car; and "the common property of Audi A4 and pickup is the only automobile. So, the semantic similarity of the former should be higher than the latter. Figure 27 shows an example of the level difference.

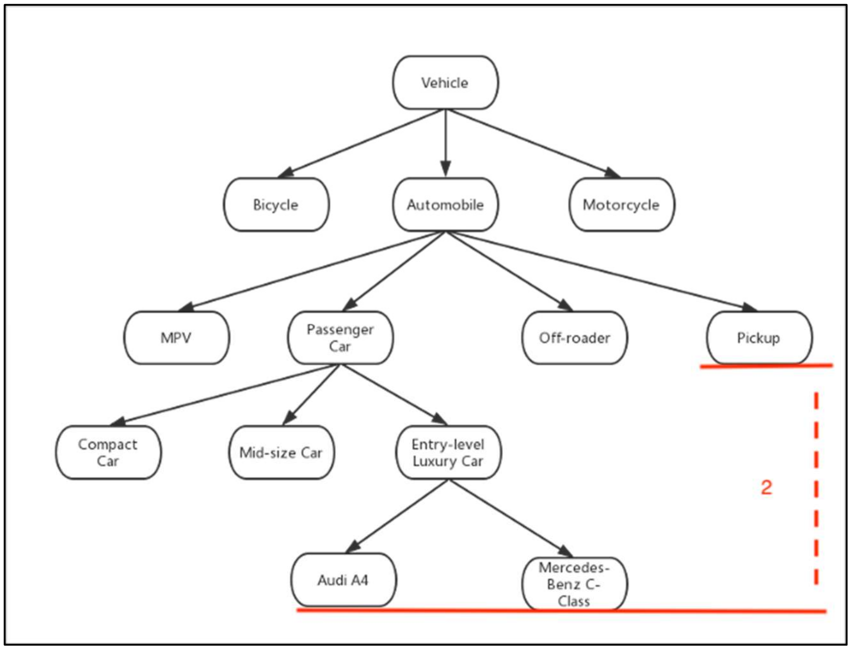


Figure 27 The example of the Level Difference

According to the concept above, we can get the Semantic Similarity between two concepts as the following formula:

$$Sim(X, Y) = \frac{\alpha \cdot \beta \cdot N(X) \cap N(Y)}{[Dis(X, Y) + \alpha] \cdot [|L(X) - L(Y)| + \beta] \cdot N(X) \cup N(Y)} \tag{7}$$

α and β are parameters, which can adjust the influence of the three factors above. We can understand that the Sim (X, Y) has a range of (0, 1], which means all the concepts are related to the ontology. Therefore, the Semantic Similarity can infinitely approach ZERO, but it cannot be ZERO; when X and Y are the same concepts, the Semantic Similarity is equal to ONE.

3.2 Knowledge Graph-based Re-ranking model

As the ontology only supports one "first-class" relation (is-a) between the concepts, this causes some limitations about representing the domain knowledge by using the ontology. The specific information in the domain knowledge is one of the crucial factors in the knowledge representation. Therefore, some domain knowledge usually needs different custom relations to describe the relations among the concepts.

Knowledge graphs such as Freebase, [26] WordNet, [27] and GeneOntology [28] have become vital resources to support many AI-related applications, such as web/mobile search, Q&A. A knowledge graph is a multi-relational graph composed of entities as nodes and relations as different types of edges. An instance of the edge is a triplet of fact (head entity, relation, and tail entity) (denoted as (h, r, t)).

This section will present a knowledge graph-based approach. The purpose of introducing the knowledge graph is that when the knowledge graph represents the domain knowledge, the system can still understand the semantic similarity among the concepts and re-scoring and re-ranking the documents based on the semantic similarity. Then, it would allow the search engine to understand users' queries more semantically and efficiently, which is the primary goal of this dissertation. Unlike the ontology-based model, we need another method to calculate the semantic similarity between nodes, because the knowledge graph is a graph structure. Thus, the knowledge graph embedding is introduced. By embedding the knowledge graph, we can map all the nodes into space according to the relationship between them. Then we can calculate all the nodes as vectors.

3.2.1 The Workflow

The framework of the knowledge graph-based approach is similar to the ontology-based approach. As the knowledge graph is a symbolic and logical system, and all the data are saved as a graph structure, thus we cannot use the same approach to calculate the semantic similarity as the ontology used. Therefore, we need to add one step-knowledge graph embedding before calculating the semantic similarity. Furthermore, we also need another approach to implement the creation of domain knowledge because the current syntax of OWL does not support the declaration and usage of new custom relations. The model includes six parts:

- Create the domain knowledge base;
- Knowledge graph embedding;
- Calculate the semantic similarity between the concepts in the knowledge graph;

- Score the candidate documents by the semantic similarity and the term's TF-IDF weight;
- Rank the candidate web pages by the score, which are generated in the previous step, and return the result.

Figure 28 shows the framework of the approach.

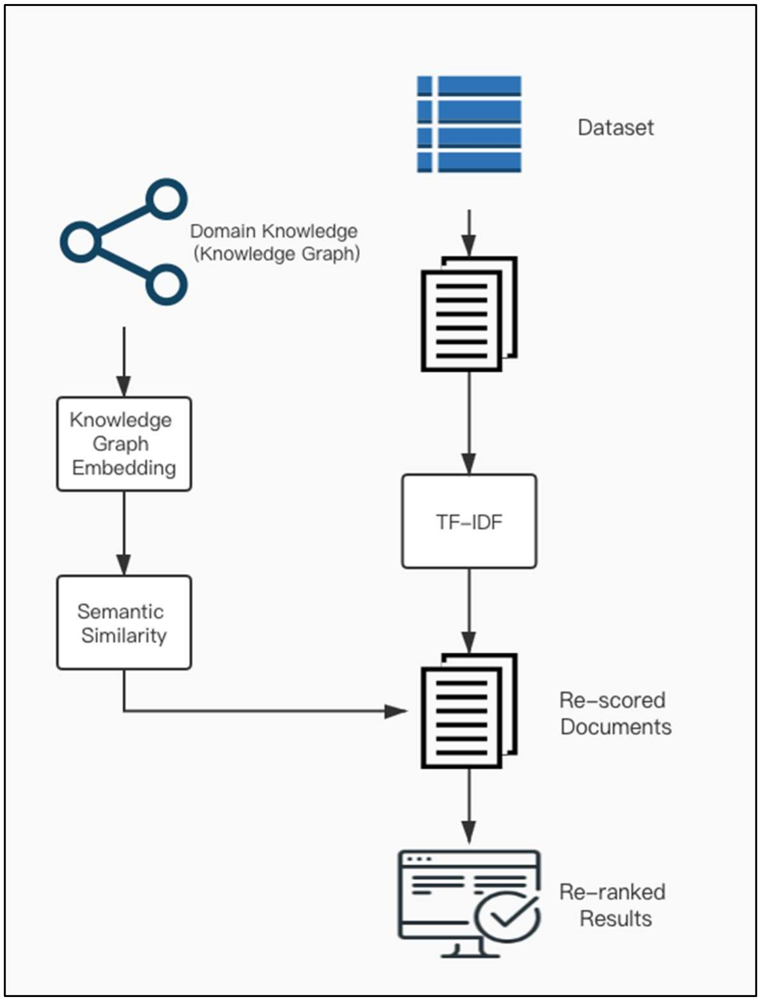


Figure 28 The framework of the knowledge graph-based model

3.2.2 Relationship in the embedding space

In order to translate a symbolic and logical system of a knowledge graph into a continuous vector space, we introduce the TransE, which is a model for learning low-dimensional embeddings of entities. The relationships among the knowledge graph can be represented as translations in the embedding space in the TransE model. [29] More specifically, if we have a triplet (*h, l, t*), then the embedding of the tail entity (denoted as *t*) should be close to the embedding of the head entity (denoted as h) plus some vector that depends on the relationship (denoted as *l*). It can be formulated as $t = h + l$. Figure 29 is an illustration of the projected vectors that are embedded by the head entity, tail entity, and the relationship.

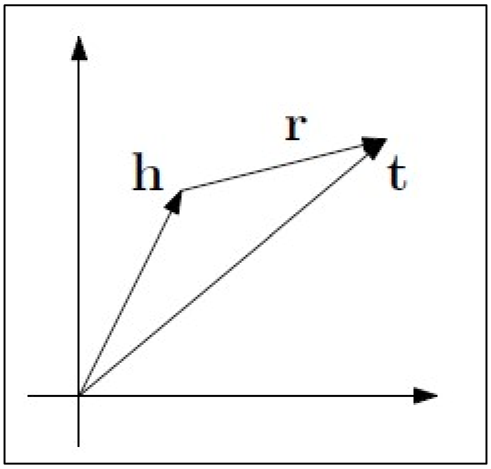


Figure 29 The illustration of TransE

According to this idea, given a knowledge graph as a training set (S) of triplets (*h, t, l*) composed of two entities $h, t \in$ E (entities set) and a relationship l ∈ L (relationships set). As the introduction previously, the basic idea behind the model is the functional relation induced by the relationship corresponds to a translation of the embeddings. For example, $h + l \approx t$ when (*h, l, t*) holds, which means the tail entity should be the nearest neighbor of the head entity plus the relationship. Moreover, the head entity plus the relationship should be far away from the tail entity otherwise.

When we try to learn the embeddings from the training set S of triplets (*h, l, t*), we just minimize the loss function:

$$\mathcal{L} = \sum\nolimits_{(h,l,t)\in S} \sum\nolimits_{(h',l,t')\in S'_{(h,l,t)}} \gamma + d(h + l, t) - d(h' + l, t')_{+} \tag{9}$$

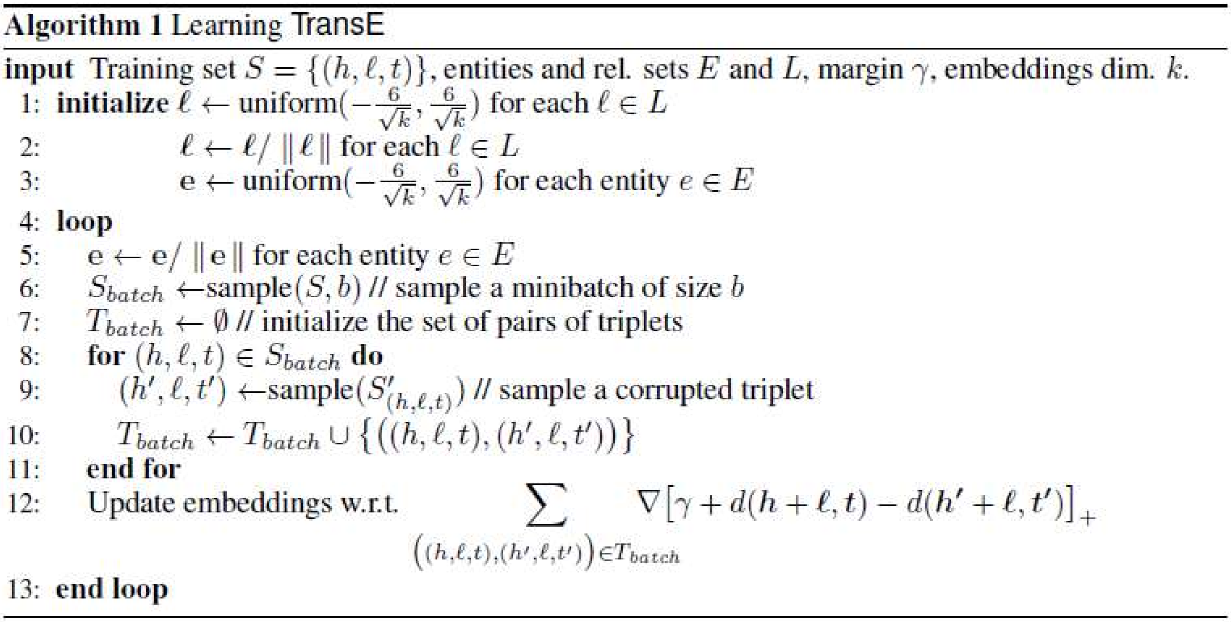
d in equation 9 is a dissimilarity measure that can be used to measure the plausibility of the triplet (h, l, t) in the embedding space. Moreover, we consider the dissimilarity measure as the squared Euclidean distance, then we have:

$$d(h + l, t) = ||h||_2^2 + ||l||_2^2 + ||t||_2^2 - 2\left(h^T t + l^T (t - h)\right) \tag{10}$$

And the $[x]_+$ in equation 5 denotes the positive part of x, $\gamma > 0$ is a margin hyperparameter, and

$$S'_{(h,l,t)} = \{(h', l, t) | h' \in E\} \cup \{(h, l, t') | t' \in E\} \tag{11}$$

The set S' is the set of corrupted triplets, which is constructed according to equation 11. The corrupted triplets are composed of training triplets with either the head entity or tail entity randomly. The detailed optimization procedure is described in Algorithm 1.



```
Algorithm 1 Learning TransE
input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets $E$ and $L$, margin $\gamma$, embeddings dim. $k$.
initialize $\ell \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$
           $\ell \leftarrow \ell / \|\ell\|$ for each $\ell \in L$
           e $\leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$
loop
    e $\leftarrow$ e$/ \|$e$\|$ for each entity $e \in E$
    $S_{batch} \leftarrow$sample$(S, b)$ // sample a minibatch of size $b$
    $T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets
    for $(h, \ell, t) \in S_{batch}$ do
        $(h', \ell, t') \leftarrow$sample$(S'_{(h,\ell,t)})$ // sample a corrupted triplet
        $T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$
    end for
    Update embeddings w.r.t. $\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla [\gamma + d(h + \ell, t) - d(h' + \ell, t')]_+$
end loop
```

3.2.3 *Knowledge embedding by Translating on Hyperplanes*

TransE performs very well for embedding a knowledge graph. However, the TransE has a flaw when dealing with mapping relationships of reflexive/one-to-many/many-to-one/many-to-many. For example, Barack Obama was the president of the United States of America, and Bill Clinton was also the president of the United States of America. Now, we have two triplets in this case, (Barack Obama, was president of, USA) and (Bill Clinton, was president of, USA). We can observe from these two triplets, both two triplets have the same relationship and the trail entity (t), and they have a different head entity (h). This case shows the potential limitations of TransE. The TransE uses the relationship as the translation to train the model; thus, the embedding vectors of these head entities will be very similar after training. The truth is evident that Barack Obama and Bill Clinton are a different entity in the real world.

Accordingly, another approach introduced here: TransH [30] that is a modified version based on the TransE. The TransH can alleviate the limitations of the TransE. The TransH interprets a relationship as a translating operation on a hyperplane. Namely, each relationship is characterized by two vectors in the TransH model, one is a normed vector (w_l), and another one is the translation vector (d_l) on the hyperplane. As the introduction mentioned before, the knowledge graph can be defined as the triplet *(h, l, t*), which is correct in terms of the real world's fact. In model TransH, we expect that the translation vector d_l connects the projections of the head entity (h) and the tail entity (t) with low error. From the illustration in Figure 30 (a), we can understand that, for a relationship (*l*), the model would be better than the embedded entities in the same space when it positions the relation-specific translation vector (d_l) in the relation-specific hyperplane (w_l, the normal vector). This approach can overcome the TransE's limitations in dealing with reflexive/one-to-many/many-to-one/many-to-many relationships while remains the model complexity almost as same as the TransE. For example, Barack Obama and Bill Clinton will be the different entities in this case, but they still can share the same relationship and the trail entity (h). As illustrated in figure 30 (b), we can assume that h represents the entity Barack Obama and h' represent the entity, Bill Clinton. h and h' are different vector, but they share the same relationship and the trail entity.

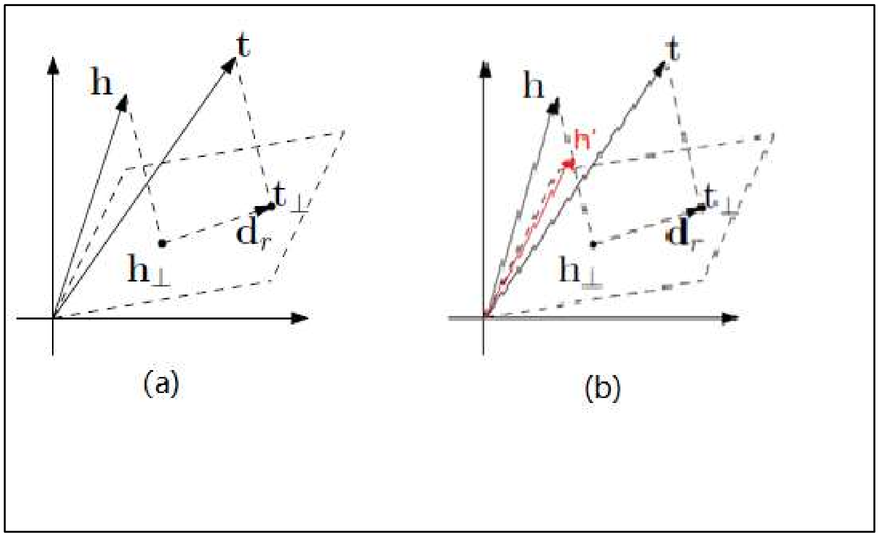


Figure 30 The illustration of TransH

Based on this concept, using the connection between the head entity's projection and the trail entity's projection as the translation, the TransH model can be described as follow.

For a triplet (*h, l, t*), first we projected the embedding h and *t* to a hyperplane w_r, and the projections are denoted as $h_\perp$ and $t_\perp$ respectively. If this triplet (*h, l, t*) is a golden triplet, we expect that the translation vector d_l which connected the projections $h_\perp$ and $t_\perp$on the hyperplane will have a low error.

The next step is to define the scoring function based on the approach listed in the first step.

$$||h_\perp + d_l - t_\perp||_2^2 \tag{12}$$

The scoring function aims to measure the plausibility that a triplet is incorrect. As the $||w_l||_2 = 1$, it is easy to obtain

$$h_\perp = h - w_l^T h w_l,$$

$$t_{\perp} = t - w_l^T t w_l \tag{13}$$

Then the score function is

$$f_l(h,t) = \left|\left|(h - w_l^T h w_l) + d_l - (t - w_l^T t w_l)\right|\right|_2^2 \tag{14}$$

If the triplet (*h, l, t*) is a golden triplet, the expectation of the score from equation 14 is lower. Otherwise, the score will be higher if the triplet (*h, l, t*) is an incorrect triplet. Once obtained the dissimilarity measure score, the loss function will be：

$$\mathcal{L} = \sum_{(h,l,t)\in\Delta} \sum_{(h',l',t')\in\Delta'_{(h,l,t)}} [f_l(h,t) + \gamma - f_{l'}(h',t')]_+ \tag{15}$$

Where the $[x]_+$ in equation 15 denotes the positive part of x, the Δ is the set of golden triplets (positive) and the Δ' denotes the set of the incorrect triplets (negative). The $\gamma > 0$ is a margin hyperparameter that is used to separate the positive and negative triplets (*h, l, t*).

As the introduction in the previous part, it is necessary to construct the negative triplets for a golden triplet during the training. In model TransE, the negative triplets are obtained by corrupting the golden triplets randomly. Namely, the model TransH has a different approach to create the negative triplets for a golden triplet. The first step is to set different probabilities for the replacing head entity (*h*) or trail entity (*t*) when corrupting the triplet. These probabilities are set that depends on the mapping property of the relationship, such as one-to-many, many-to-one, or many-to-many. By using this approach, it can obtain a false negative label. More specifically, we denote the average number of tail entities (t) of head entities (h) as tph; and the average number of head entities (h) per tail entities (t) as *h, p* and *t* respectively. Then we can define a Bernoulli distribution with the parameter $\frac{tph}{tph+h}$ for sampling. For example, if we have a golden triplet (*h, l, t*) of the relationship *l*, we corrupt the triplet by replacing the head entity (h) with the probability $\frac{tph}{tph+hpt}$, and corrupt the triplet by replacing the tail entity (t) with the probability $\frac{hpt}{tph+hpt}$.

3.2.4 Semantic Similarity

As mentioned before, in order to obtain the relevant web pages, semantic similarity can be used to measure the relevance between the query and the other concepts. For instance, in the ontology-based model, the semantic similarity is calculated by Semantic Distance, Semantic Coincidence, and Level Difference based on the tree-structured model. And then using the semantic similarity as a weight added on the keyword-based scoring function. At last, the model will return the candidate web pages in the descending order where the score comes from the scoring function with the semantic similarity weight. This approach has an excellent performance in the ontology-based model because it has a tree structure. However, it cannot be used as a method for calculating the semantic similarity in a knowledge graph. Because the concepts in knowledge are different. In order to solve this problem, the Cosine Similarity will be introduced below.

Cosine Similarity is a measurement of similarity between two non-zero vectors of inner product space, and the similarity is measured based on the cosine of the angle between them. And these two non-zero vectors represent two terms to measure their similarity. For example, as figure 31 (a) shows, two vectors v1 and v2 represent two terms, respectively. The cosine of the angle between these two vectors represents the similarity of the two terms (Cosθ). Two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° related to each other have a similarity of 0, and two diametrically opposed vectors have a similarity of -1, independent of their magnitude. The cosine similarity is mainly used in positive space, where the outcome is neatly bounded in [0,1]. Namely, the angle between the two vectors is smaller, which means that the cosine value is closer to 1; thus, the corresponding terms are more similar. In figure 31 (b), the similarity between v1 and v2 is more significant than v2 and v3.

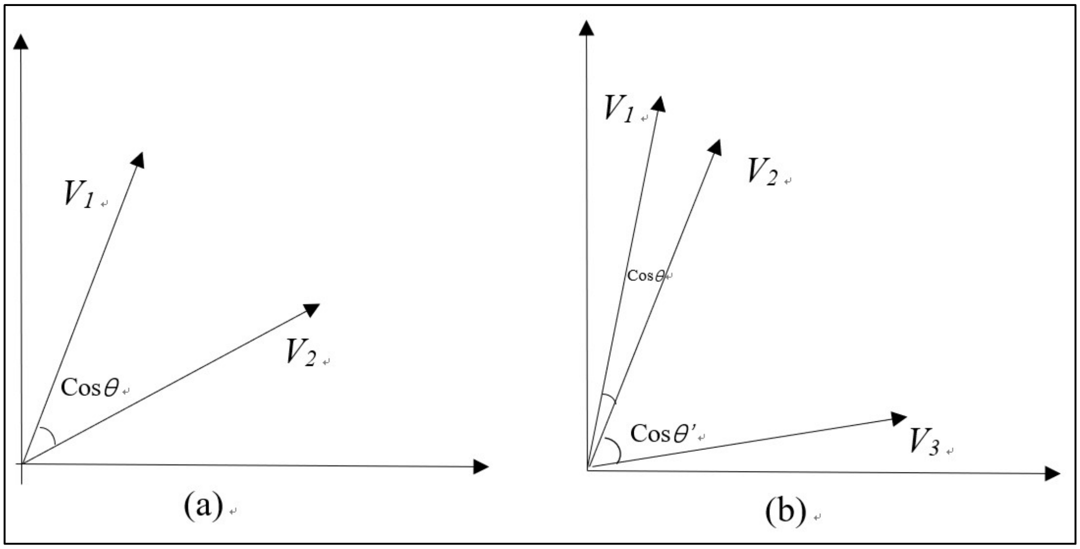


Figure 31 Cosine Similarity

In a triangle, the cosine value of an angle can be calculated by the following formula:

$$\cos(\theta) = \frac{a^2 + b^2 - c^2}{2ab} \tag{16}$$

Moreover, the cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\mathrm{A} \cdot \mathrm{B} = \left|\left|\mathrm{A}\right|\right|\left|\left|\mathrm{B}\right|\right| \cos\theta \tag{17}$$

Therefore, the cosine similarity $(\cos\theta)$ can be obtained by the given two vectors of attributes, which is represented by using a dot product and magnitude as

$$Simlarity = \cos\theta = \frac{\mathrm{A} \cdot \mathrm{B}}{\left|\left|\mathrm{A}\right|\right|\left|\left|\mathrm{B}\right|\right|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i}\sqrt{\sum_{i=1}^{n} B_i}} \tag{18}$$

3.3 The candidate documents re-ranking

In the previous section, two approaches were introduced to covert a knowledge (ontology and knowledge graph) into a machine-readable quantification result. When we have solved the problem that the symbolic and logical system (domain knowledge) involves numerical computing in continuous spaces, we can get the semantic similarity among concepts in domain knowledge.

If all concepts in domain knowledge are treated as a dictionary, the query keywords (assuming the keywords belong to the concept in the domain knowledge), and all candidate documents could be found in the dictionary and the candidate documents. Then, the re-ranking algorithm will score each candidate's documents by using the semantic similarity and the term's TF-IDF weight.

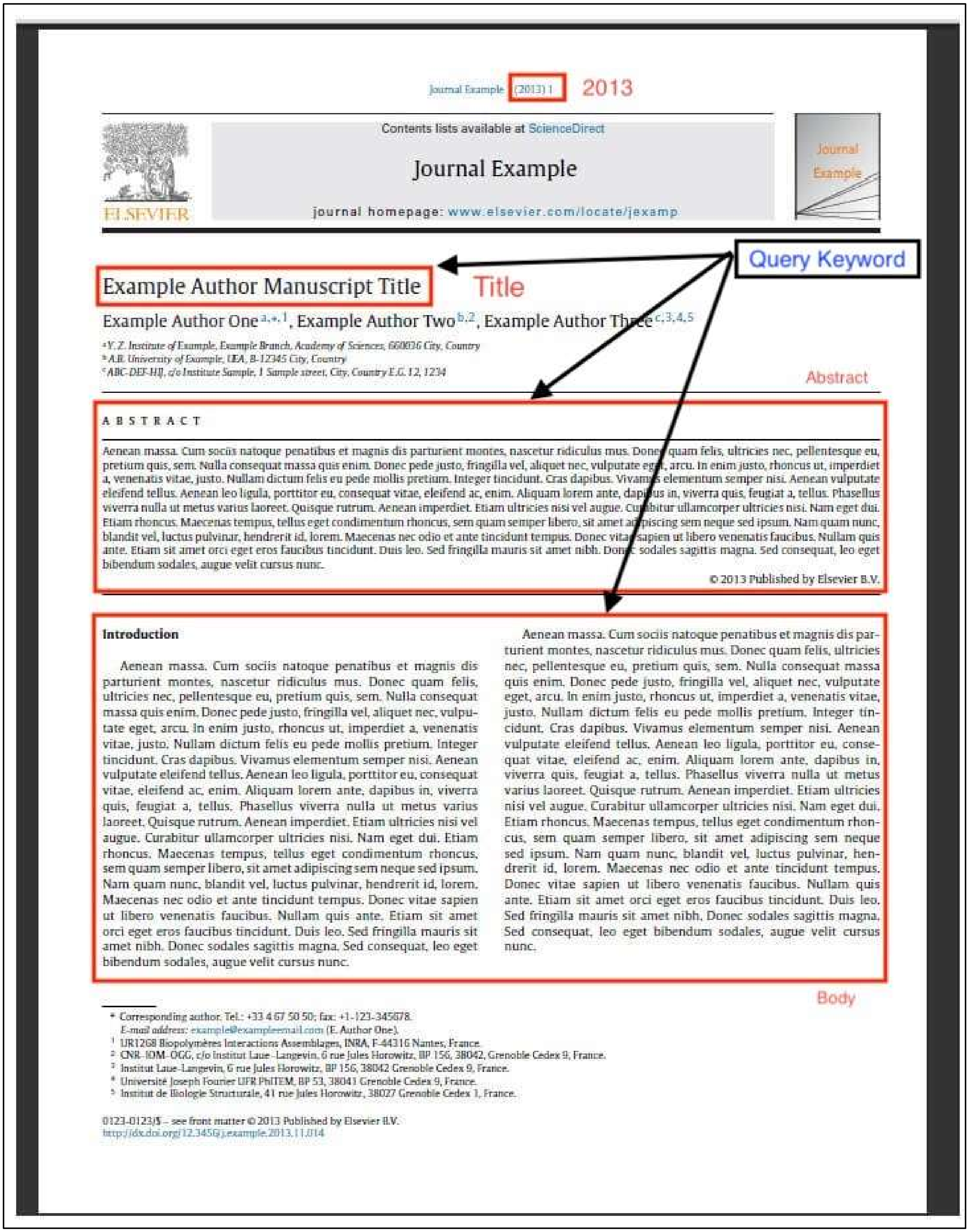
As a result, a score function can be represented as follows:

$$Score(keyword) = TFIDF(keyword) + \sum_{i}[TFIDF(wi) \times Sim(keyword, wi)] \quad (19)$$

The wi means the words in the knowledge base except for the query keyword. Moreover, *Sim (A, B)* refers to the semantic similarity between *A* and *B*.

In addition to scoring each candidate document based on the TF-IDF value of the query keyword and the semantic similarity within the domain knowledge, the improved methods proposed in this dissertation also considers the location of the query keyword and related concepts among the candidate documents. That is, when the query keywords and related concepts appear in different places in the document (such as title, abstract, keywords), the weight of the document will be different. For example, if a query keyword appears in the title, the weight given to the document is higher than the weight obtained when the keyword appears only in the body. At the same time, the improved method proposed in this dissertation also considered that if multiple candidate documents have the same relevance to the search keywords, the latest published documents will get higher weight.

Figure 32 is a demonstration of an article. As we described in the previous paragraph, a search keyword appears in different places in the article, and the associated weights will be different. If multiple candidate documents have the same score, the rank of them will be judged by their publication time. For example, this article was published in 2013. If an article was published in 2018, we consider that the article published in 2018 has a higher score.



Journal Example (2013) 1

Contents lists available at ScienceDirect

Journal Example

journal homepage: www.elsevier.com/locate/jexamp

ELSEVIER

Journal Example

Example Author Manuscript Title

Example Author One[a,*,1], Example Author Two[b,2], Example Author Three[c,3,4,5]

[a] *Y. Z. Institute of Example, Example Branch, Academy of Sciences, 660036 City, Country*
[b] *A.B. University of Example, UEA, B-12345 City, Country*
[c] *ABC-DEF-HIJ, c/o Institute Sample, 1 Sample street, City, Country E.G. 12, 1234*

ABSTRACT

Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

© 2013 Published by Elsevier B.V.

Introduction

Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc.

[*] Corresponding author. Tel.: +33 4 67 50 50; fax: +1-123-345678.
E-mail address: example@exampleemail.com (E. Author One).

[1] UR1268 Biopolymères Interactions Assemblages, INRA, F-44316 Nantes, France.

[2] CNR-IOM-OGG, c/o Institut Laue-Langevin, 6 rue Jules Horowitz, BP 156, 38042, Grenoble Cedex 9, France.

[3] Institut Laue-Langevin, 6 rue Jules Horowitz, BP 156, 38042 Grenoble Cedex 9, France.

[4] Université Joseph Fourier UFR PhITEM, BP 53, 38041 Grenoble Cedex 9, France.

[5] Institut de Biologie Structurale, 41 rue Jules Horowitz, 38027 Grenoble Cedex 1, France.

0123-0123/$ – see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/12.3456/j.example.2013.11.014

Figure 32 A demonstration of an article

After each document has been re-scored, the next task is to sort each document from highest to lowest according to the score of each document; namely, the highest-scoring document represents that it is most relevant to the query keyword.

At last, the re-ranked documents from the dataset will be returned to the users.

3.4 Recommendation system based on the domain knowledge

After the search engine has acquired the ability to calculate semantic similarity from the ontology and the knowledge graph, domain knowledge can be used to simulate semantic search. So far, the search engine has a similar knowledge background as domain experts to optimize search. In order to further improve the user's search experience, this dissertation will propose a recommendation system based on domain knowledge.

After the knowledge base is introduced into the search engine, the search engine will obtain all the concept nodes in the domain knowledge. The search engine can narrow down the search scope according to the introduced domain knowledge, and recommend these possible search scopes to users, in order to help users to achieve more accurate searches. For example, when the search keyword is the automobile, according to the introduced knowledge base, the search engine will obtain relevant knowledge. That is, the automobile includes MPV, passenger car, off-roader, and pickup. According to this knowledge, in order to help users to narrow down the search scope, search engines can divide the search into different search areas, such as MPV, passenger car, off-roader, or pickup. Of course, users can reject these suggestions and still perform general searches.

The goal of the recommendation system is to help users to complete the search more accurately. Moreover, the system will conduct the recommendation in the following steps (figure 33 is shown the workflow of the recommendation system)..

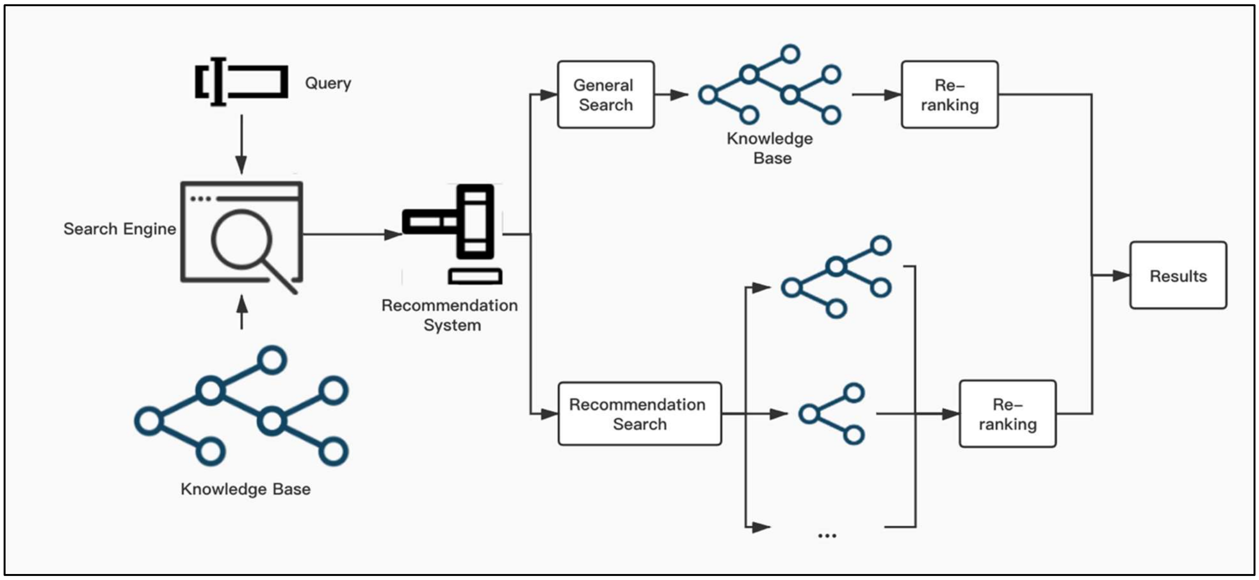


Figure 33 The workflow of the recommendation system

First, when the query keyword is entered into the search engine, the recommendation system will give users the following options: The first option is the general search. Users do not need the recommendations offered by the system. They will select the input query as the keyword to feed into the knowledge base. Then, the system will use the knowledge base and combine the semantic similarity calculation method proposed in this dissertation to score the candidate documents and return the results to the users based on the descending order of the score; The second option is recommendation search. The recommendation system will return the possible precise retrieve range to the users based on the domain knowledge. For example, if the query keyword is the automobile, according to the introduced knowledge base, the search engine will obtain relevant knowledge; that is, the automobile includes MPV, passenger car, off-roader, and pickup. Then, the system will switch sub-domain based on the user selected. The sub-domain means the system will switch the query keyword into one of the recommended concepts (figure 34 is a demonstration of the sub-domain knowledge from the recommendation system) and combine the semantic similarity calculation method proposed in this dissertation to score the candidate documents. According to the calculation method of semantic similarity described in the previous sections, once the query keywords changed, the semantic similarity between concepts in the knowledge base changes as well. Namely, all concepts will be recalculated based on this new keyword to achieve the primary purpose- narrowing down the search range and finding more semantic relate documents with the new keywords.

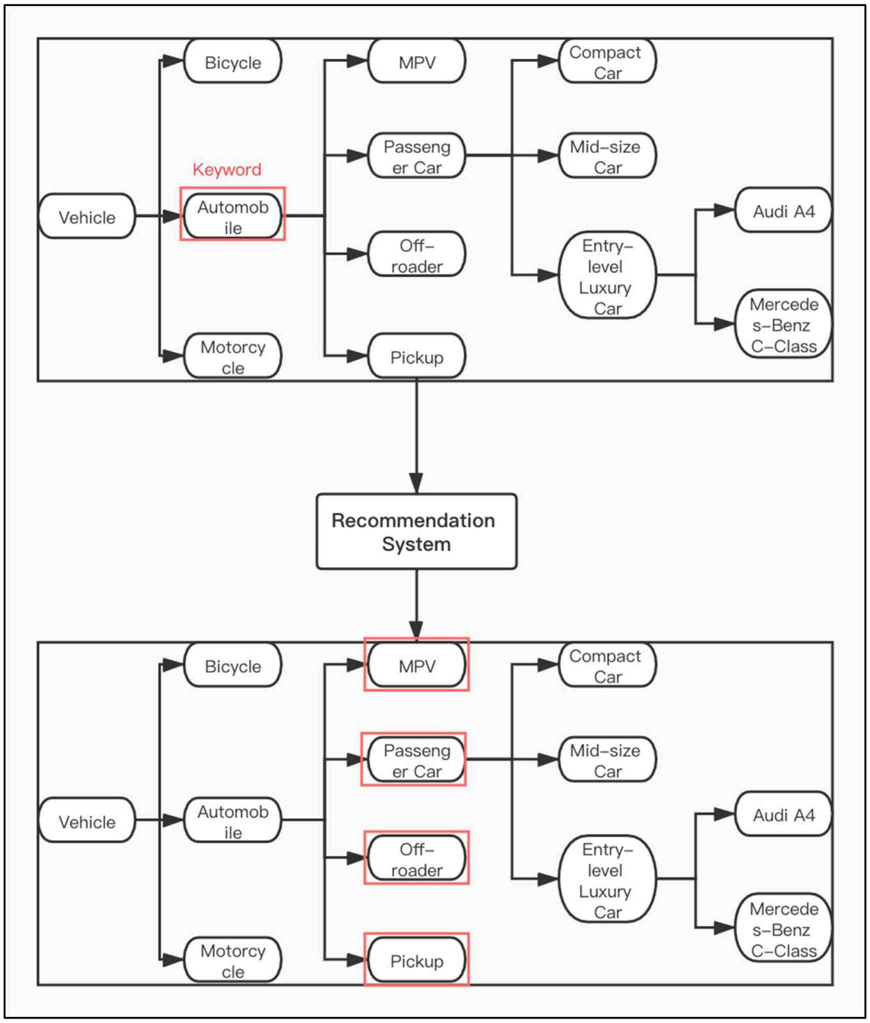


Figure 34 Demonstration of sub-domain knowledge from the recommendation system

By using the proposed method, the users can use the recommendation system to narrow the search to a more precise range automatically.

3.5 Conclusion

In this chapter, first, two methods for acquiring concepts from domain knowledge and calculating the semantic similarity among concepts are proposed. In order to simulate the semantic search in the existing search engine, it is essential to let the search engine to understand the introduced domain knowledge. Namely, the search engine should understand the semantic relationship among the concepts in the domain knowledge. The ontology can express the domain knowledge and the knowledge graph, according to the different structure of the knowledge base (ontology and knowledge graph), I proposed two methods to calculate the semantic similarity among the concepts in the domain knowledge. Then, the search engine re-scores each document by combining the TF-IDF weights and the position of the query keyword appeared in the document. At last, the search engine returned the re-ranked results based on the scores. This chapter also proposed a recommendation system based on the introduced domain knowledge. The recommendation system can narrow down the search scope according to the introduced domain knowledge and recommend these possible search scopes to users to help users complete more accurate searches than the general search. According to these proposed methods, we can improve the users' search experience by letting the search engine achieve the capabilities to capture the conceptualizations involved in users' intention and meaning of web pages' content step by step.

Chapter 4 Experiment

In order to verify the proposed ideas in the previous chapters can be attainable to the desired goals, this chapter will sketch the structure and implement an experimental prototype system.

4.1 Introduction

In the previous chapters, we have already discussed the existing solutions in the current search engine. The ranking algorithms such as PageRank, HITS can be used to improve the accuracy of retrieval for the search engine. In order to improve the retrieval, more and more methods have been introduced into the search engine. However, all the algorithms and methods we discussed need to combine the keyword-based algorithm to optimize the ranking. That means the traditional ranking algorithm in the current search engine needs to rely on the frequency of keywords to determine the degree of relevance between the query and the documents or web pages. For example, the PageRank algorithm can filter the web pages' quality based on the number of links. If there are many links link to a web page, we can consider that web page has high quality, if there is a high-quality web page link to another one, then the linked web page can be considered as a high-quality web page. With this idea, those low-quality pages can be filtered out. However, it still requires using the keywords to determine the relevance of the webpage to the query. In most instances, if the query keyword appears the most in the document or a web page, the traditional ranking algorithm will determine that a document or a web page has the highest degree of relevance to the query.

The characteristics of the traditional ranking algorithms through this dissertation research listed below:

- The traditional ranking algorithm needs to combine the keyword-based algorithm to determine the relevance between the query and the documents or web pages.
- The optimal ranking algorithm introduced in previous chapters can only help search engines filter out low-quality webpages and find out high-quality webpages.

Based on these characteristics of traditional ranking algorithms, some limitations still exist in current search engines. For example, if the queries have abbreviations or similar terms, the traditional search engines will not recognize them. For example, if a user wants to search “iPhone” in a search engine, although a web page mentioned about “Apple is trying to make a new type of mobile phone,” traditional search engines are powerless in this situation because they based on literal matching as the basis for sorting. The result of no literal matching (no keyword “iPhone” on the web page) will not be searched, the query and the web page are semantically relevant.

The semantic search was introduced to solve these problems. By using metadata, the search engine can be “semantic.” More specifically, the resource objects in the network will be annotated into the search engine (the metadata of each document or web pages will be embedded). Then, the search engine can process the users’ query expression semantically. However, annotating every web page leads to inefficiency for the search engine. Since a domain knowledge can be described by an ontology or knowledge graph, the search engine will be able to calculate the semantic similarity among the concepts in domain knowledge and add the similarity as a weight in the keyword-based search. As a result, the search engine will have the capabilities to capture the conceptualizations involved in users' intention and web pages' content meaning.

4.2 Experiment Design

The purpose of this dissertation is to find a more efficient way for search engines to complete semantic queries in a specific domain. An experiment should be designed to verify the proposed methods' performance, whether they can improve the current traditional keyword-based search and emulate the effects of semantic search or not.

In this experiment, the experiment result expects to prove that the proposed methods can provide more accurate query results that match the users' real intention than the current existing search engine's retrieval result. In order to achieve this goal, the results should satisfy the following assumptions:

- Some documents are highly semantic related with the query but ranked very behind by using the existing traditional search engine.
- When the knowledge base was introduced (ontology-based model or knowledge graph-based model), the search engine can rank the highly semantic related documents to the top.

Based on the assumptions, the designed experiment should be completed as the following steps:

- Select an existing search engine platform

- Introduce a domain knowledge
- Design a prototype for re-ranking models
- Design a recommendation system that is proposed in chapter 3.
- Compare the different return lists from the existing search engine platform and the re-ranking models

To implement the designed experiment needs to match three criteria: 1) a professional information retrieval platform 2) in a particular subject field 3) proposed re-ranking methods in this dissertation. First, the experiment will obtain the search results (ranked) for a particular query keyword returned by the platform. Second, domain knowledge will be designed by the expert in a specific subject field. Then the results are re-ranked by the methods proposed in this dissertation. In order to compare the performance between the existing search engine platform and the re-ranking models, the volunteers who are composed of the experts in the field will be involved in the experiment, and they will make a judgment on the list of results obtained by different methods. By analyzing their professional judgment results, the performance of different methods (existing method on the platform, and the methods proposed in this dissertation) can be determined. A questionnaire will be designed for volunteers from the same subject field. Since this experiment involves human participants, I have completed the CITI Program (Collaborative Institutional Training Initiative)'s training and obtained the certificate. Since this experiment is not a human subjects research, there is no need to obtain the Pace University IRB approval. The designed questionnaire will contain the following sections:

- Different return lists from existing search engine platforms and re-ranking models. However, the lists are not identified for participants.
- The participants will rate different lists (include the recommended lists which are returned by the designed recommendation system) to rate the lists by using the Likert-scale survey.

4.2.1 Data Preparation

The experiment expects to verify that the proposed methods can understand users' queries more semantically and effectively in a specific domain than the traditional search engine. Once the search engine obtains a domain knowledge, it should reuse the knowledge flexibly through the methods proposed in this dissertation. Therefore, the experiment should focus on a specific field, and the data should be diverse in order to compare the performance of different models effectively.

ERIC (Education Resources Information Center) is an internet-based digital library of education research and information sponsored by the Institute of Education Sciences (IES) of the U.S. Department of Education. ERIC provides access to bibliographic records of journal and non-journal literature from 1966 to the present. Furthermore, over half a million users search the ERIC website each week, with many more searching through ERIC data using vendor sites. ERIC users include education researchers, students, teachers, librarians, administrators, education policymakers, parents, and the general public.

The ERIC collection includes bibliographic records (citations, abstracts, and other pertinent data) for 1.6 million items since1966, including:

- Journal articles
- Books
- Research syntheses
- Conference papers
- Technical reports
- Policy papers and
- Other education-related materials

Currently, over 1,000 journals are indexed in ERIC. Almost all of these journals are indexed comprehensively. Every single article in each issue and each volume has been included in ERIC. A small number are indexed selectively. Only those articles that are education-related are selected for indexing. Besides, authors and publishers have given ERIC permission to display more than 350,000 full-text materials available with no charge. Many of these materials are "grey literature" such as conference papers and reports, but there are a growing number of journal articles and books freely available in ERIC as well. Most materials published in 2004 and forward included links to other sources, including publishers' websites.

As we introduced previously, education can be seen as a specific domain for this experiment. ERIC is a professional platform to provide extensive literature for the public. At the same time, ERIC uses a keyword-based search when retrieving documents for users. Therefore, the experiment will be tested in the field of education by using the ERIC platform, and the experiment will compare the result lists from the ERIC and re-ranked by the proposed methods in the dissertation, thereby confirming the hypothesis proposed by the experiment.

ERIC returned a total of 22,359 kinds of literature by using the query keyword “exploration.” The experiment extracted the titles, abstracts, publication time, and description keywords of these documents according to the order returned by ERIC. Based on the extracted information, the documents were established. The experiment fed these documents to the re-ranking methods to generate the re-ranked result lists for the next step.

4.2.2 The domain knowledge

As the designed experiment has set the search field of education, the experiment invited some experts in the field of education to participate in the design of the knowledge base. The experts consist of Ph.D. students and master students in the Education department at one top private university in the US. After identifying the search keywords for the experiment, we determined the scope covered by the knowledge base according to the relevant knowledge of the concept.

The Community of Inquiry (CoI) framework theory, methodology, and instruments were developed during a Canadian Social Sciences and Humanities research funded project entitled “A Study of the Characteristics and Qualities of Text-Based Computer Conferencing for Educational Purposes” project which ran from 1997 to 2001. Central to the original study was the creation of a model of a community of inquiry comprised of three essential elements of an educational experience:

- Cognitive presence

- Social presence
- Teaching presence

Outcomes of the original project were published in peer-reviewed journals, which, in turn, have resulted in hundreds of research studies applying and extending the original CoI theory, method, and instruments. The seminal paper "Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education" [31] has been cited more than 5,605 times (as reported by Google Scholar December 2019) and provided the foundation for valuable empirical research in learning theory across multiple disciplines and varied educational settings.

This experiment invited the experts in the field of education to design a domain knowledge based on the CoI theory. The concept "community of inquiry" is the kernel term of the theory. The community of inquiry comprised of three essential elements of an educational experience: cognitive presence, social presence, and teaching presence. The basic idea of the domain knowledge is that the different presences are described more concretely when moving away from the kernel. For example, cognitive presence means that the students can feel that their thinking is related to the teaching content. Moreover, the classifications of the cognitive presence create perspectives or start actions, explore, integrates, come up with solutions. So, the terms triggering event, exploration, integration, and resolution are part of the term cognitive presence.

With the experiment's purpose, the knowledge base has been designed into two versions. The first version contains a total of 14 concepts that are related to the community of inquiry, and the second version contains a total of 53 concepts. The overall architecture of two designed knowledge bases are shown in figures 35 and 36.

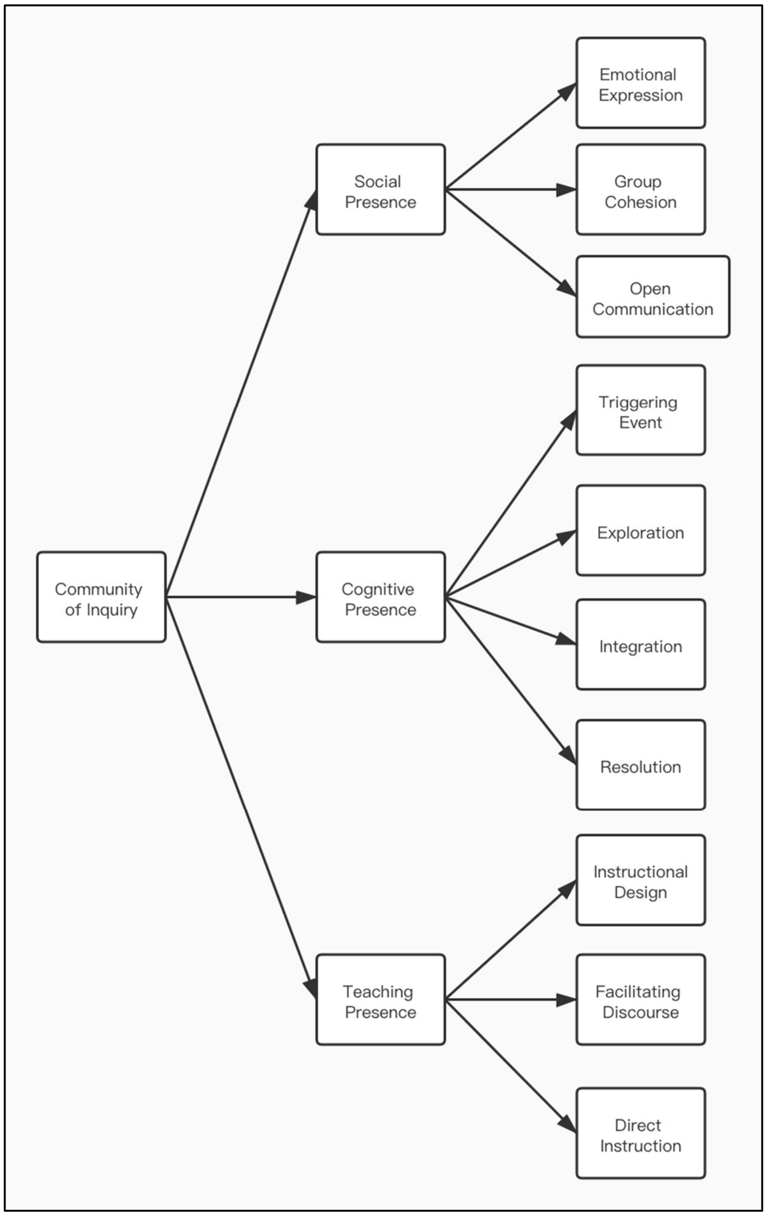


Figure 35 The overall architecture of the first knowledge base

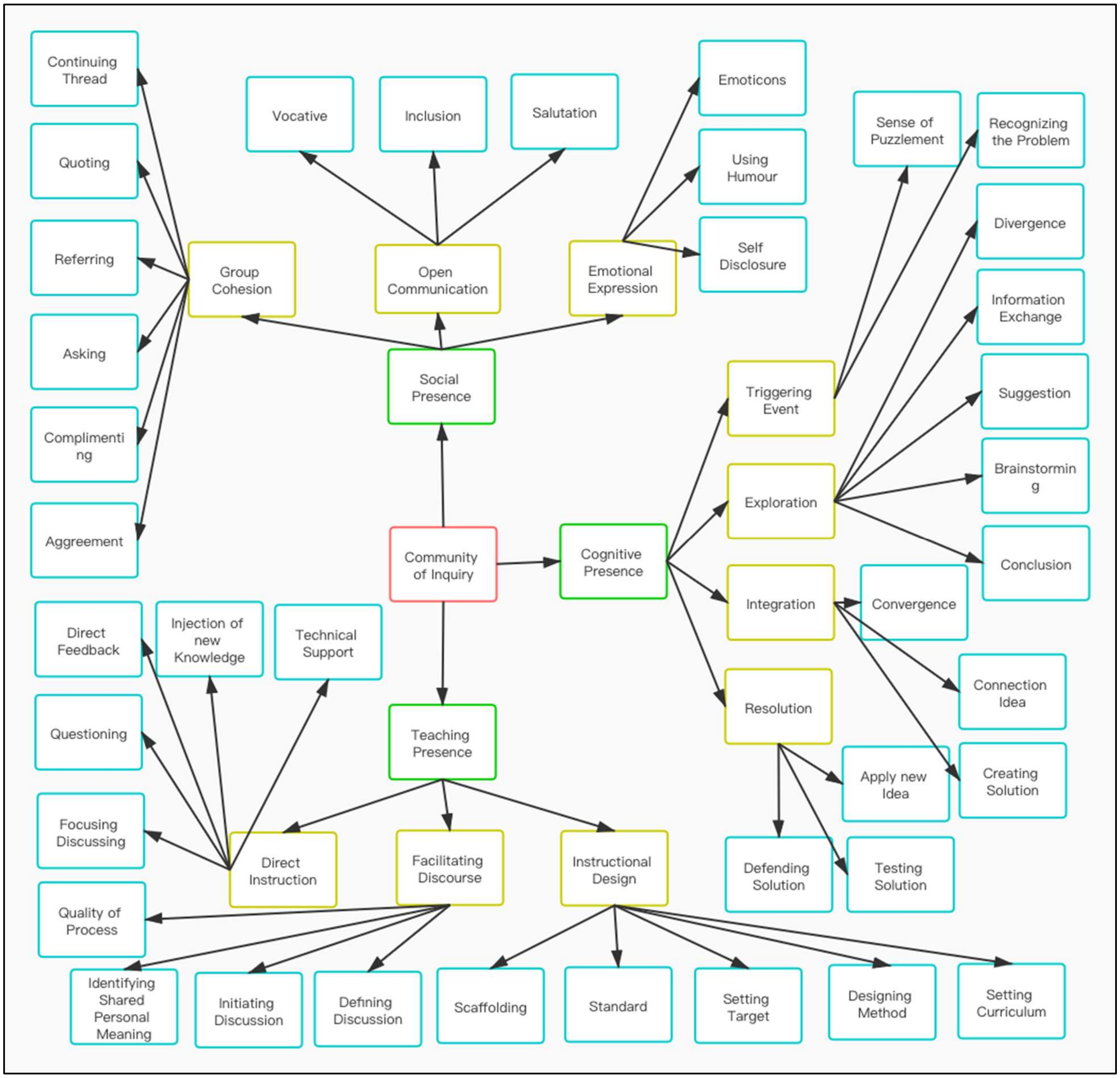


Figure 36 The overall architecture of the second knowledge base

The traditional search engines optimize the results ranking by the algorithms such as PageRank, HITS algorithm. However, these ranking algorithms can only help search engines filter out low-quality webpages and find out high-quality webpages. The search engines still need to determine the degree of association between the query and the documents on the query keyword appears frequency in the webpages. Therefore, the re-ranking methods proposed in this dissertation will be more semantic and effective by introducing domain knowledge.

According to the content of the knowledge base, the experiment also uses cognitive presence, triggering event, integration, and resolution as keywords to retrieve on the ERIC platform, and extracts their retrieval results as a data set for the implementation of the recommendation system designed in this dissertation.

4.2.3 A prototype for re-ranking models

By designing and applying the re-ranking algorithms, the experiment will feed the results returned by the ERIC into the designed re-ranking models in order to achieve the following assumptions:

1. The problem should be solved when the knowledge base is introduced. Furthermore, the more comprehensive the knowledge is expressed, the more semantically related documents will be retrieved.
2. The documents that are highly relevant to the query but have a low keyword frequency should be ranked ahead when the ontology-based and knowledge-graph based ranking algorithm applied in the experimental prototype. Because the traditional keyword-based ranking algorithm only considers the frequency of the keyword, while evaluating the degree of relevance between the query and the document. After domain knowledge was introduced, the proposed re-ranking algorithm would evaluate the degree of relevance between the query and the document that not only depends on the keyword

but, more importantly, it also considers whether the document can better match the domain knowledge.

4.2.4 *The related work for the prototype design*

The designed prototype will implement re-ranking algorithms. Moreover, the prototype is built with JavaScript, HTML5, and Cascading Style Sheets using Django as the web framework. As the introduction in the previous section, the query result from the ERIC platform will be fed into the re-ranking algorithm as the dataset, and then re-ranked results will be shown as the new query results. So, the web application of the designed prototype will demonstrate three query results. The query result returned from the ERIC, and the re-ranked query results returned based on the different knowledge bases, respectively. The workflow of the designed prototype is shown in figure 37.

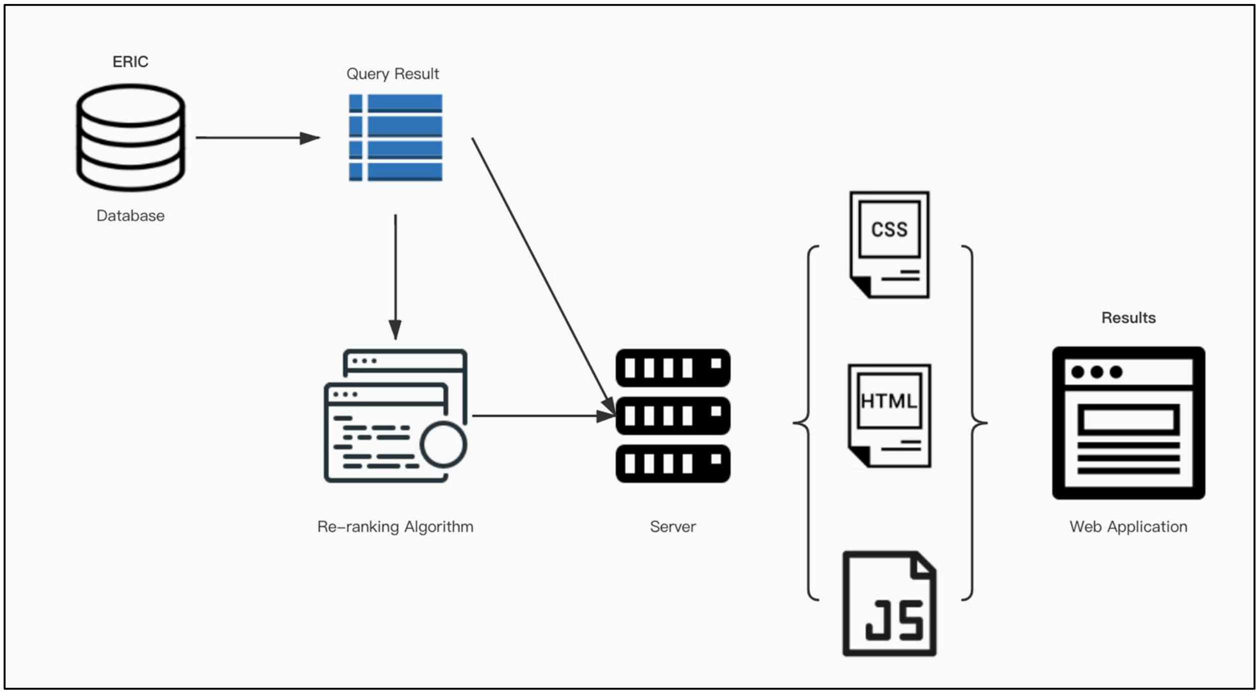


Figure 37 The workflow of the prototype

Django (https://www.djangoproject.com/) is a Python-based free and open-source web framework, which follows the model-template-view (MTV) architectural pattern.

Hypertext Markup Language (HTML) is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript.

Cascading Style Sheets (CSS) is a style sheet language to describe the presentation of a document written in an HTML.

JavaScript (https://www.javascript.com/about), often abbreviated as JS, is most well-known as the scripting language for web pages.

Alongside HTML and CSS, JavaScript is one of the core technologies of the World Wide Web. JavaScript enables interactive web pages and is an essential part of web applications. The vast majority of websites use it for client-side page behavior, and many major web browsers have a dedicated JavaScript engine to execute it.

After the Django server is set up, the re-ranked query results will be displayed in a web application by using technologies such as HTML, CSS, and JavaScript. The prototype was developed with the assistance of the PyCharm, which is an integrated development environment (IDE) used in computer programming, specifically for the Python language.

4.3 The details of the re-ranking prototyping

In order to compare the proposed methodologies in this dissertation and the existing solution in the traditional search engine (ERIC, for this experiment), the re-ranking prototype will be developed and conducted by the following steps:

- Prepare the experiment dataset (see the detail in 4.1).
- Build a prototype with the proposed methods in this dissertation.
- Feed the results returned by the ERIC and let the prototype score each document.
- Re-rank the results generated by the prototype.
- Implement the experimental prototype on a Graphical User Interface

The workflow of the prototype with the existing solution is shown in figure 38.

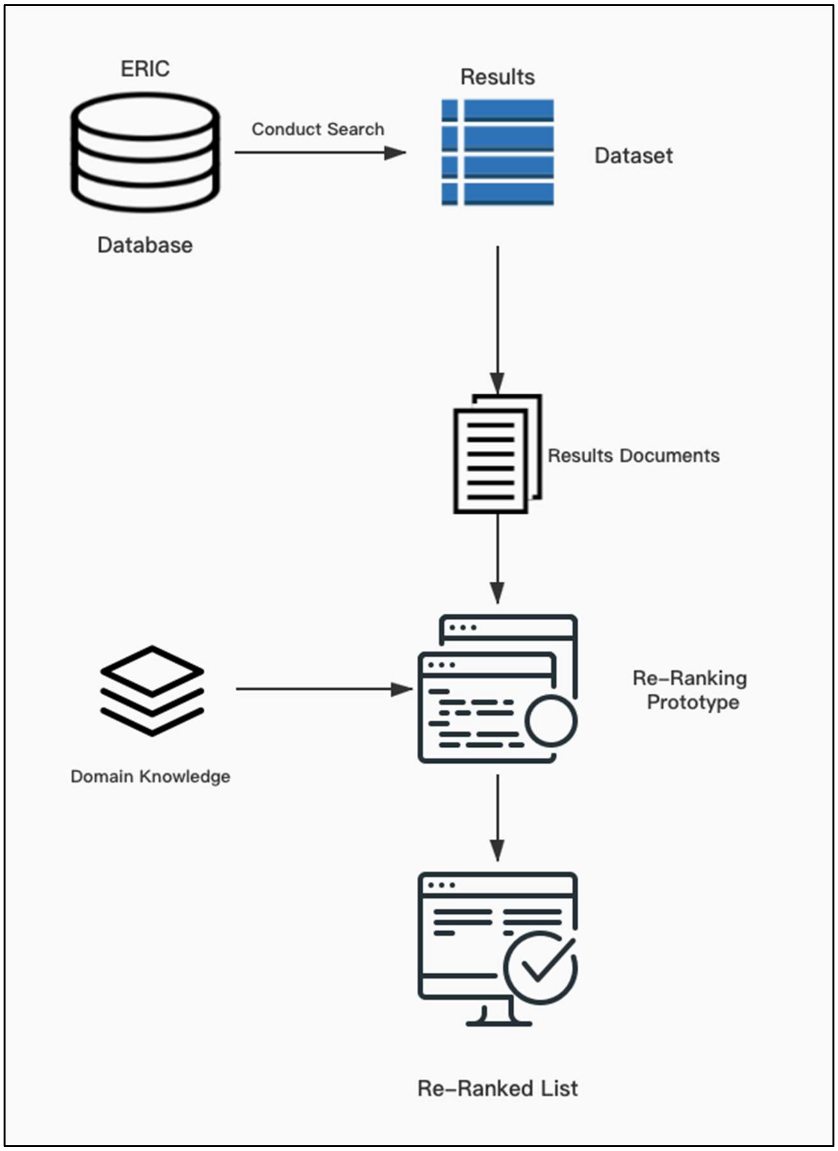
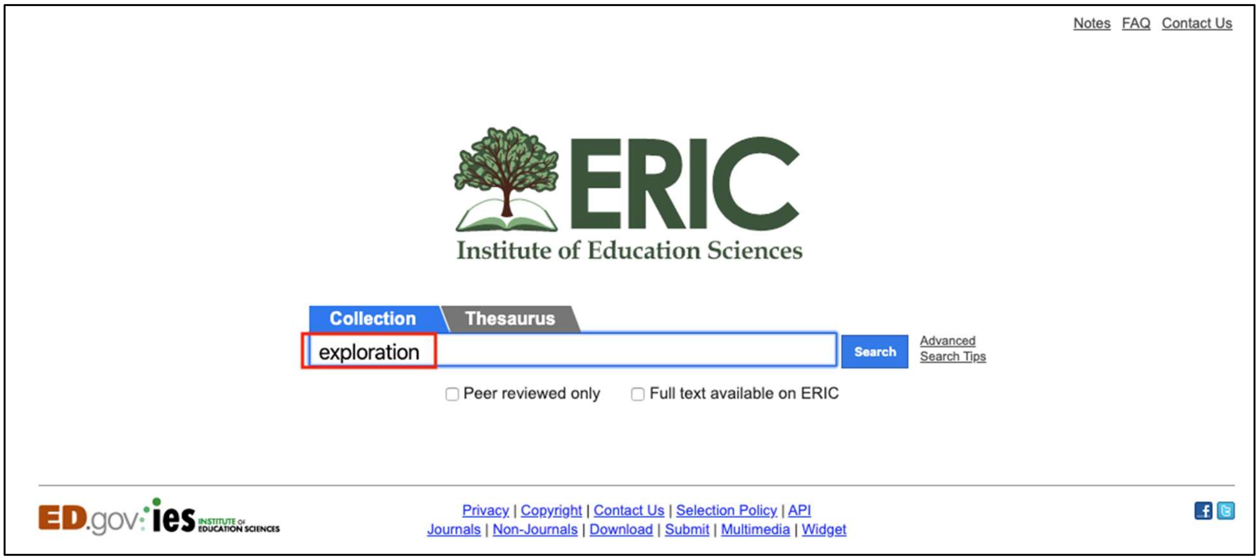


Figure 38 The workflow of the re-ranking prototype

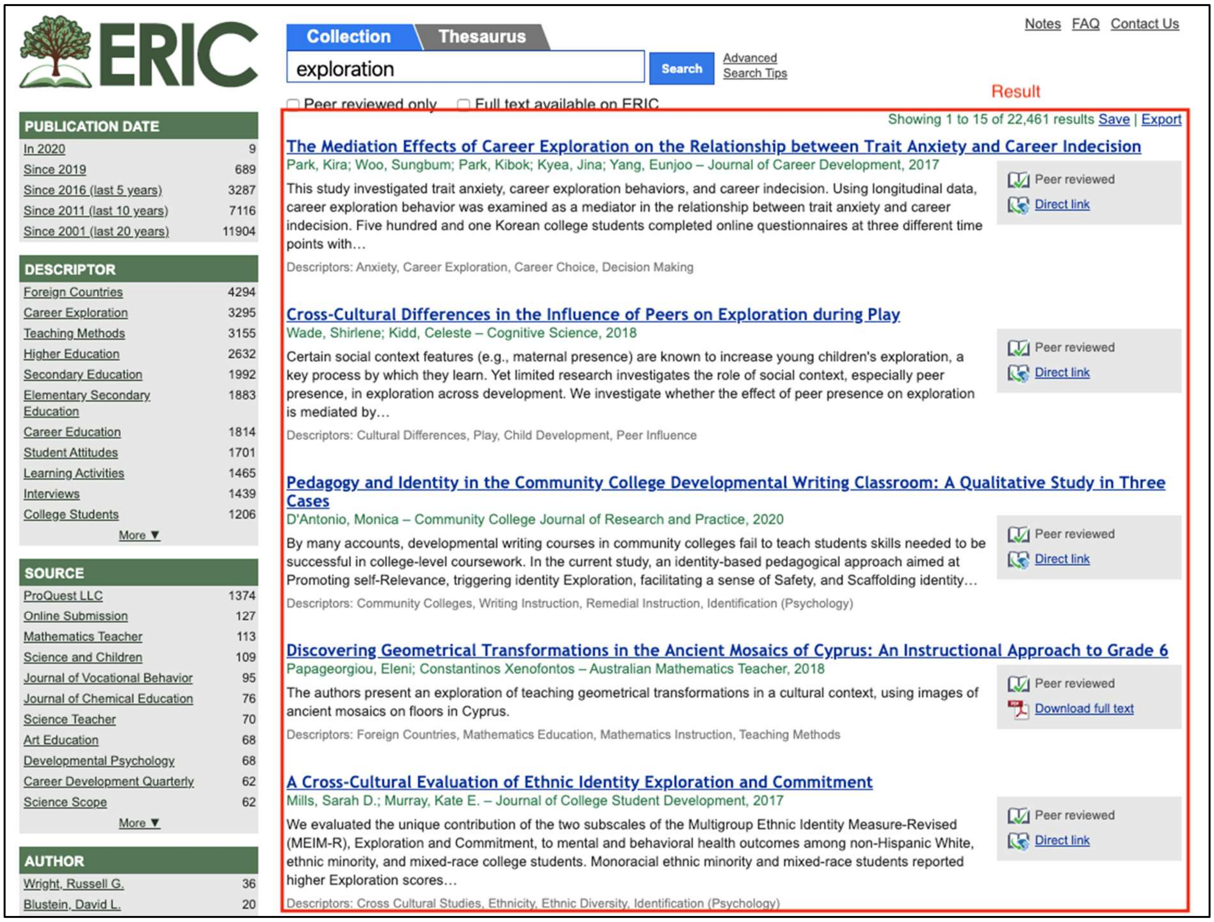
First, we will prepare the experimental data set for the experiment. Candidate documents come from the ERIC (https://eric.ed.gov/) platform. The demonstration set the query keyword as “exploration” and use it to search on the ERIC platform (see figure 39)..


Notes FAQ Contact Us
ERIC
Institute of Education Sciences
Collection
Thesaurus
Search
Advanced
Search Tips
□ Peer reviewed only
□ Full text available on ERIC
ED.gov ies INSTITUTE OF EDUCATION SCIENCES
Privacy | Copyright | Contact Us | Selection Policy | API
Journals | Non-Journals | Download | Submit | Multimedia | Widget

exploration

Figure 39 Query in ERIC

According to the retrieval results returned by ERIC, the prototype exported and processed the results (figure 40).



exploration

PUBLICATION DATE	
In 2020	9
Since 2019	689
Since 2016 (last 5 years)	3287
Since 2011 (last 10 years)	7116
Since 2001 (last 20 years)	11904

DESCRIPTOR	
Foreign Countries	4294
Career Exploration	3295
Teaching Methods	3155
Higher Education	2632
Secondary Education	1992
Elementary Secondary Education	1883
Career Education	1814
Student Attitudes	1701
Learning Activities	1465
Interviews	1439
College Students	1206

SOURCE	
ProQuest LLC	1374
Online Submission	127
Mathematics Teacher	113
Science and Children	109
Journal of Vocational Behavior	95
Journal of Chemical Education	76
Science Teacher	70
Art Education	68
Developmental Psychology	68
Career Development Quarterly	62
Science Scope	62

AUTHOR	
Wright, Russell G.	36
Blustein, David L.	20

Showing 1 to 15 of 22,461 results Save | Export

The Mediation Effects of Career Exploration on the Relationship between Trait Anxiety and Career Indecision
Park, Kira; Woo, Sungbum; Park, Kibok; Kyea, Jina; Yang, Eunjoo – Journal of Career Development, 2017

This study investigated trait anxiety, career exploration behaviors, and career indecision. Using longitudinal data, career exploration behavior was examined as a mediator in the relationship between trait anxiety and career indecision. Five hundred and one Korean college students completed online questionnaires at three different time points with…

Descriptors: Anxiety, Career Exploration, Career Choice, Decision Making

Peer reviewed
Direct link

Cross-Cultural Differences in the Influence of Peers on Exploration during Play
Wade, Shirlene; Kidd, Celeste – Cognitive Science, 2018

Certain social context features (e.g., maternal presence) are known to increase young children's exploration, a key process by which they learn. Yet limited research investigates the role of social context, especially peer presence, in exploration across development. We investigate whether the effect of peer presence on exploration is mediated by…

Descriptors: Cultural Differences, Play, Child Development, Peer Influence

Peer reviewed
Direct link

Pedagogy and Identity in the Community College Developmental Writing Classroom: A Qualitative Study in Three Cases
D'Antonio, Monica – Community College Journal of Research and Practice, 2020

By many accounts, developmental writing courses in community colleges fail to teach students skills needed to be successful in college-level coursework. In the current study, an identity-based pedagogical approach aimed at Promoting self-Relevance, triggering identity Exploration, facilitating a sense of Safety, and Scaffolding identity…

Descriptors: Community Colleges, Writing Instruction, Remedial Instruction, Identification (Psychology)

Peer reviewed
Direct link

Discovering Geometrical Transformations in the Ancient Mosaics of Cyprus: An Instructional Approach to Grade 6
Papageorgiou, Eleni; Constantinos Xenofontos – Australian Mathematics Teacher, 2018

The authors present an exploration of teaching geometrical transformations in a cultural context, using images of ancient mosaics on floors in Cyprus.

Descriptors: Foreign Countries, Mathematics Education, Mathematics Instruction, Teaching Methods

Peer reviewed
Download full text

A Cross-Cultural Evaluation of Ethnic Identity Exploration and Commitment
Mills, Sarah D.; Murray, Kate E. – Journal of College Student Development, 2017

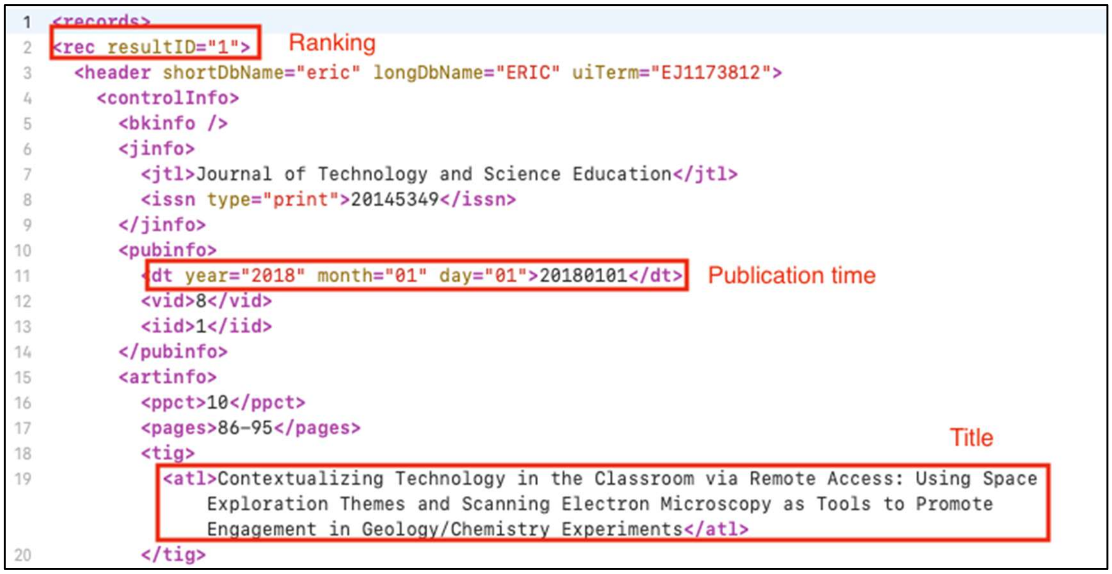
We evaluated the unique contribution of the two subscales of the Multigroup Ethnic Identity Measure-Revised (MEIM-R), Exploration and Commitment, to mental and behavioral health outcomes among non-Hispanic White, ethnic minority, and mixed-race college students. Monoracial ethnic minority and mixed-race students reported higher Exploration scores…

Descriptors: Cross Cultural Studies, Ethnicity, Ethnic Diversity, Identification (Psychology)

Peer reviewed
Direct link

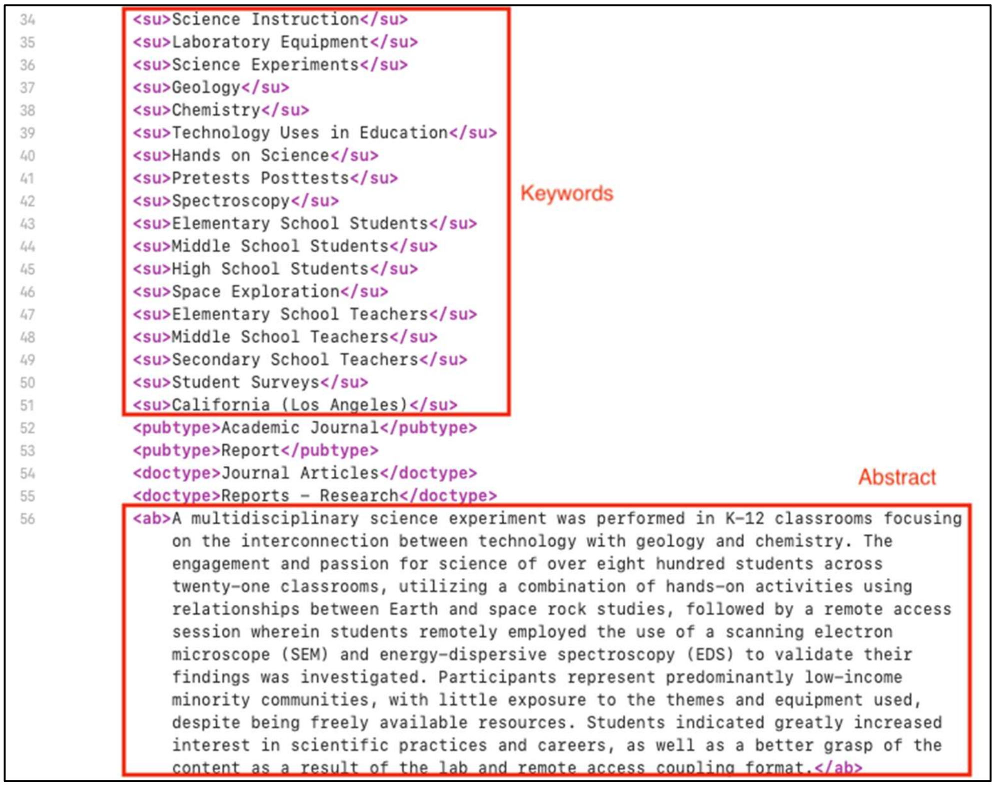
Figure 40 The query results

The query results extracted from ERIC are saved in the XML file format. According to the re-ranking method proposed in Chapter 3, the experiment extracted the title, abstract, and keywords of the article, and the publication time of the article. Then, the prototype fed this extracted information into the designed re-ranking prototype and recommendation system. As shown in figures 41 and 42, the tag "rec" represents the ranking of the article in the ERIC search results, the tag "dt" represents the time of publication, and the tag "atl" represents the title of the article. The XML file uses the tag "su" to indicate the keywords of the article and the tag "ab" to indicate the abstract of the article.



```
<records>
<rec resultID="1">
  <header shortDbName="eric" longDbName="ERIC" uiTerm="EJ1173812">
    <controlInfo>
      <bkinfo />
      <jinfo>
        <jtl>Journal of Technology and Science Education</jtl>
        <issn type="print">20145349</issn>
      </jinfo>
      <pubinfo>
        <dt year="2018" month="01" day="01">20180101</dt>
        <vid>8</vid>
        <iid>1</iid>
      </pubinfo>
      <artinfo>
        <ppct>10</ppct>
        <pages>86-95</pages>
        <tig>
          <atl>Contextualizing Technology in the Classroom via Remote Access: Using Space
              Exploration Themes and Scanning Electron Microscopy as Tools to Promote
              Engagement in Geology/Chemistry Experiments</atl>
        </tig>
```

Figure 41 The exported result from ERIC (a)



```
<su>Science Instruction</su>
<su>Laboratory Equipment</su>
<su>Science Experiments</su>
<su>Geology</su>
<su>Chemistry</su>
<su>Technology Uses in Education</su>
<su>Hands on Science</su>
<su>Pretests Posttests</su>
<su>Spectroscopy</su>
<su>Elementary School Students</su>
<su>Middle School Students</su>
<su>High School Students</su>
<su>Space Exploration</su>
<su>Elementary School Teachers</su>
<su>Middle School Teachers</su>
<su>Secondary School Teachers</su>
<su>Student Surveys</su>
<su>California (Los Angeles)</su>
<pubtype>Academic Journal</pubtype>
<pubtype>Report</pubtype>
<doctype>Journal Articles</doctype>
<doctype>Reports - Research</doctype>
<ab>A multidisciplinary science experiment was performed in K-12 classrooms focusing
    on the interconnection between technology with geology and chemistry. The
    engagement and passion for science of over eight hundred students across
    twenty-one classrooms, utilizing a combination of hands-on activities using
    relationships between Earth and space rock studies, followed by a remote access
    session wherein students remotely employed the use of a scanning electron
    microscope (SEM) and energy-dispersive spectroscopy (EDS) to validate their
    findings was investigated. Participants represent predominantly low-income
    minority communities, with little exposure to the themes and equipment used,
    despite being freely available resources. Students indicated greatly increased
    interest in scientific practices and careers, as well as a better grasp of the
    content as a result of the lab and remote access coupling format.</ab>
```

Figure 42 The exported result from ERIC (b)

After the dataset for the experiment has been prepared, the experimental prototype starts to score each document in the data set based on the introduced knowledge base.

- First, the experimental prototype calculated the term frequency and inversed document frequency of the query keywords in a document. *TF-IDF (k)* is used to refer to the term frequency and inverse document frequency of the keywords.
- Second, each document has been weighted according to the introduced knowledge base. This weight was calculated in two parts, semantic weight and the published year. For the semantic weight, 1. The experimental prototype calculated the semantic similarity between the query keywords and other concepts in the knowledge base. *Sim (k, t)* is used to refer to the semantic similarity between the query keyword and the concepts in the knowledge base. 2. The experimental prototype calculated the term frequency and inversed document frequency of the concepts in the knowledge base (except for query keywords). *TF-IDF (t*) is used to refer to them. Then, the weight of the document can be formed as:

$$semantic\ weight = \sum_i [TF - IDF(t_i) \times Sim(k, t_i)] \tag{20}$$

In the equation 20, the t_i means the concept in the knowledge base except for the query keyword.

- The experimental prototype added the corresponding weight to the document according to the time of publication. The more recent the article is published, the higher weight the experimental prototype will be assigned to the document.

- Then, the score of the document in the data set can be calculated by the term frequency and inverse document frequency of the query keywords, semantic weight, and the published year of the article. Moreover, the score of a document can be formed as:

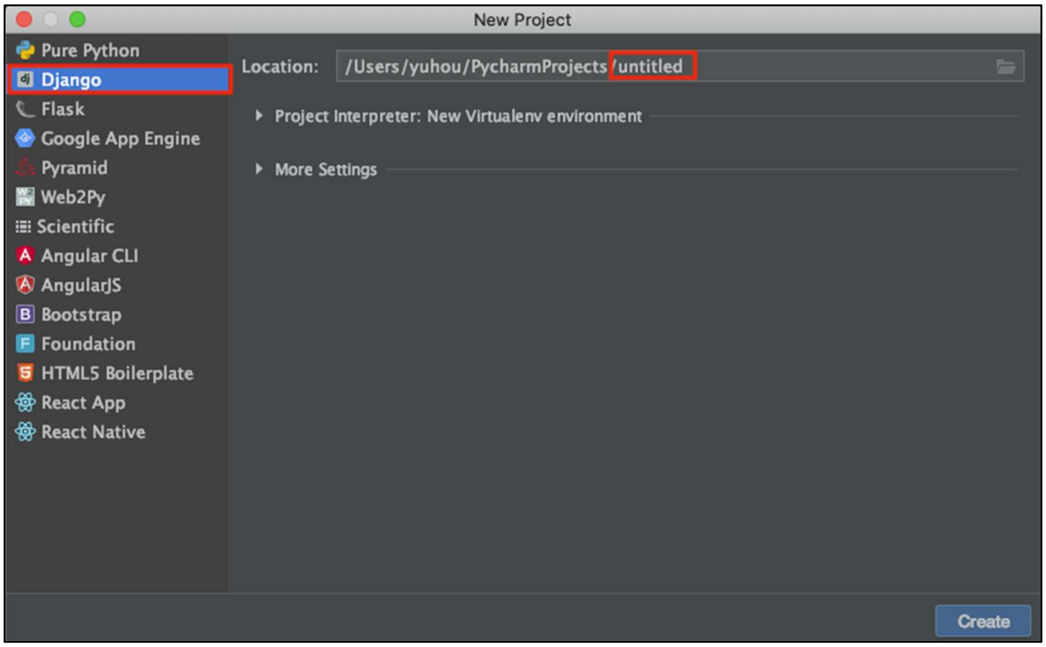
$$score_{document} = TF - IDF(k) + semantic\ weight + published\ year \qquad (21)$$

- When each document in the data set received a score from an experimental prototype, the experimental prototype starts to sort the documents based on these scores. The documents are sorted according to the score from high to low. The highest scoring document is considered to be the most relevant for retrieval.

At last, the prototype obtained different re-ranked retrieval results based on the different knowledge bases.

The experimental prototype also provides recommendations for user retrieval based on the introduced knowledge base. When a user enters a query keyword that they want to retrieve, the experimental prototype will select the concepts directly connected to the keyword as a recommendation. Then the experimental prototype will re-select new query keywords based on the user's selection(s) and calculate the score for each document.

For the next, the experiment created a Graphical User Interface to demonstrate the experiment. The GUI is built with JavaScript, HTML5, and Cascading Style Sheets using Django as the web framework. The PyCharm is used to help develop the GUI of the experimental prototype. When python3, Django, and PyCharm are installed, the first task is to create a Django project. Figure 43 shows the creation of a new Django project using PyCharm.



Location: /Users/yuhou/PycharmProjects/untitled

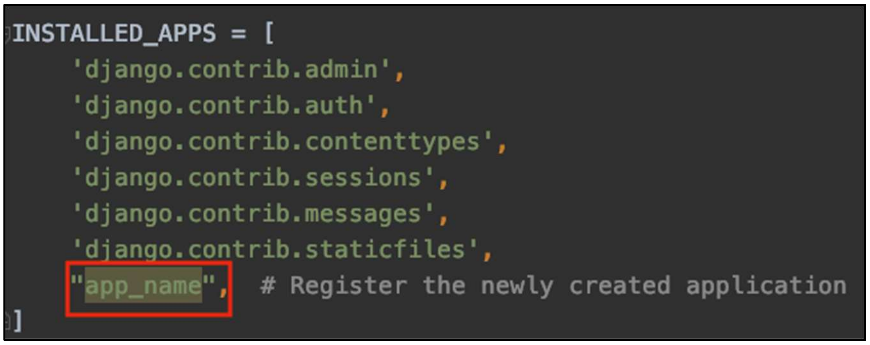
Project Interpreter: New Virtualenv environment

Figure 43 The Creation of a Django project via PyCharm

The next task is to create an application for the experimental prototype in a new Django project. The following command was used to create an application, in the same directory as manage.py:

```
$ python manage.py startapp app_name
```

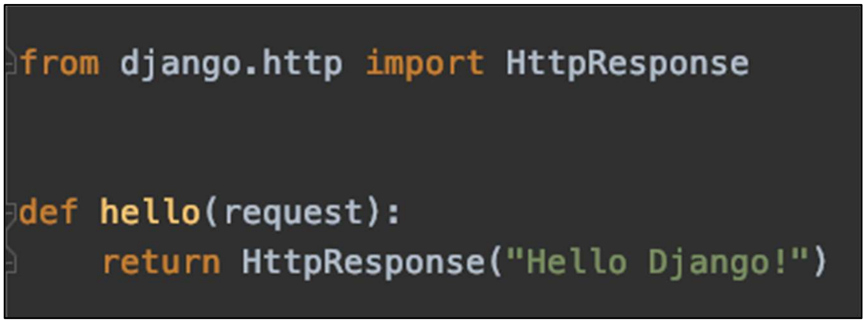
Moreover, registered in the INSTALLED_APPS in the settings.py file to ensure that the program can find this application (as shown in figure 44).



```
INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.sessions',
    'django.contrib.messages',
    'django.contrib.staticfiles',
    "app_name",  # Register the newly created application
]
```

Figure 44 The registration of the application in setting.py

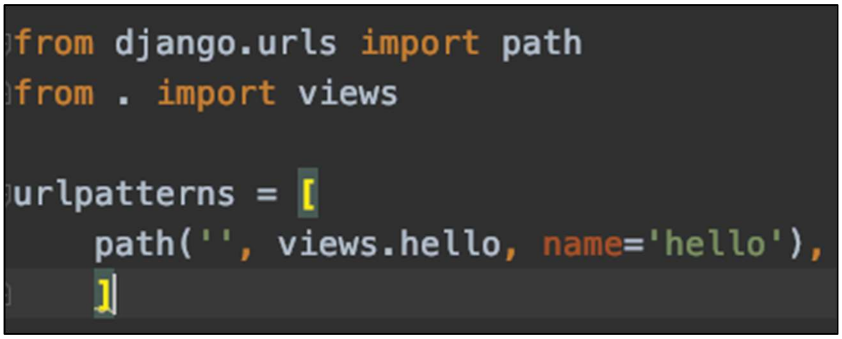
The following task is to configure the view.py file. As shown in figure 45, the Django created a new view.py file in the directory of the previously created application and entered the code:



```
from django.http import HttpResponse


def hello(request):
    return HttpResponse("Hello Django!")
```

Figure 45 Configure the view.py

The next step is to bind the URL to the view function; and open url.py and configure the settings in the file, as shown in figure 46.



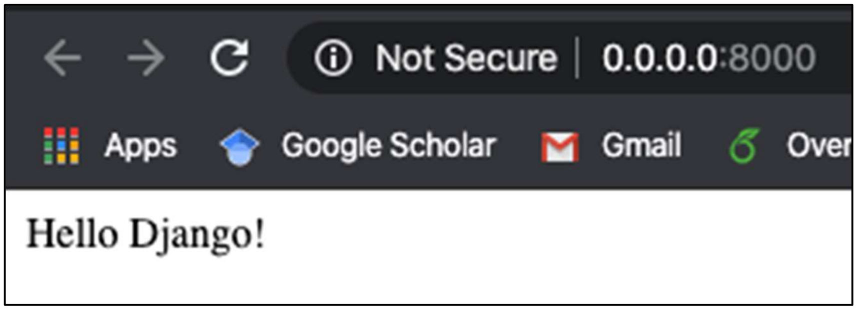
```
from django.urls import path
from . import views

urlpatterns = [
    path('', views.hello, name='hello'),
]
```

Figure 46 Configure the url.py

After the above configuration completed, the following command is used to start the Django server:

```
$ python manage.py runserver
```

The default port number is 8000. When the address: http://localhost: 8000 is entered in the browser, we can see the following interface (as shown in figure 47). Now, the Django server was successfully set up.



Hello Django!

Figure 47 Start the Django server

Next, the HTML, CSS, and JavaScript are used to design the interface of the experimental prototype. The interface of the experimental prototype is mainly composed of two parts, which are re-rank the query result from the ERIC and search with the recommendation system. Figure 48 shows the selected options in the designed prototype's interface.

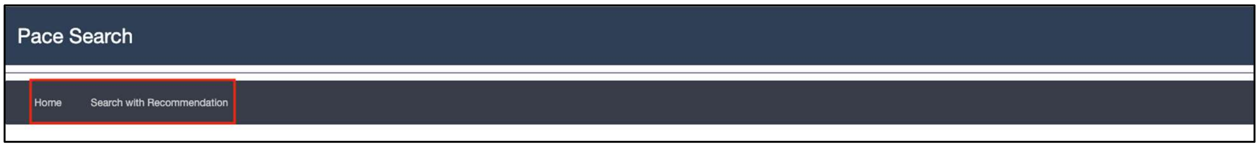
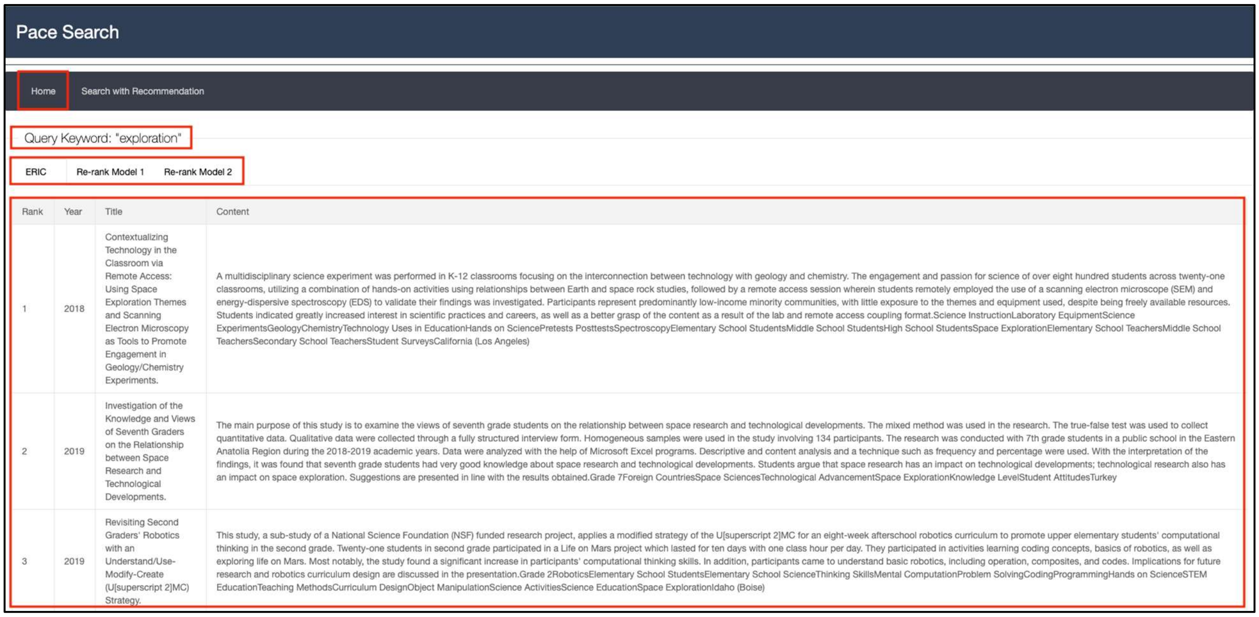


Figure 48 The selected options in prototype's interface.

In the re-rank section, the interface shows the query keywords, query results returned by ERIC, and query results returned by the proposed re-ranking to the users. Because the prototype is only used to show the implementation of the re-rank method proposed in this dissertation, therefore, the prototype will show the retrieval results with the 'exploration' as the keyword from the ERIC and re-rank methods proposed in this dissertation. Figure 49 is the interface of the re-rank section.



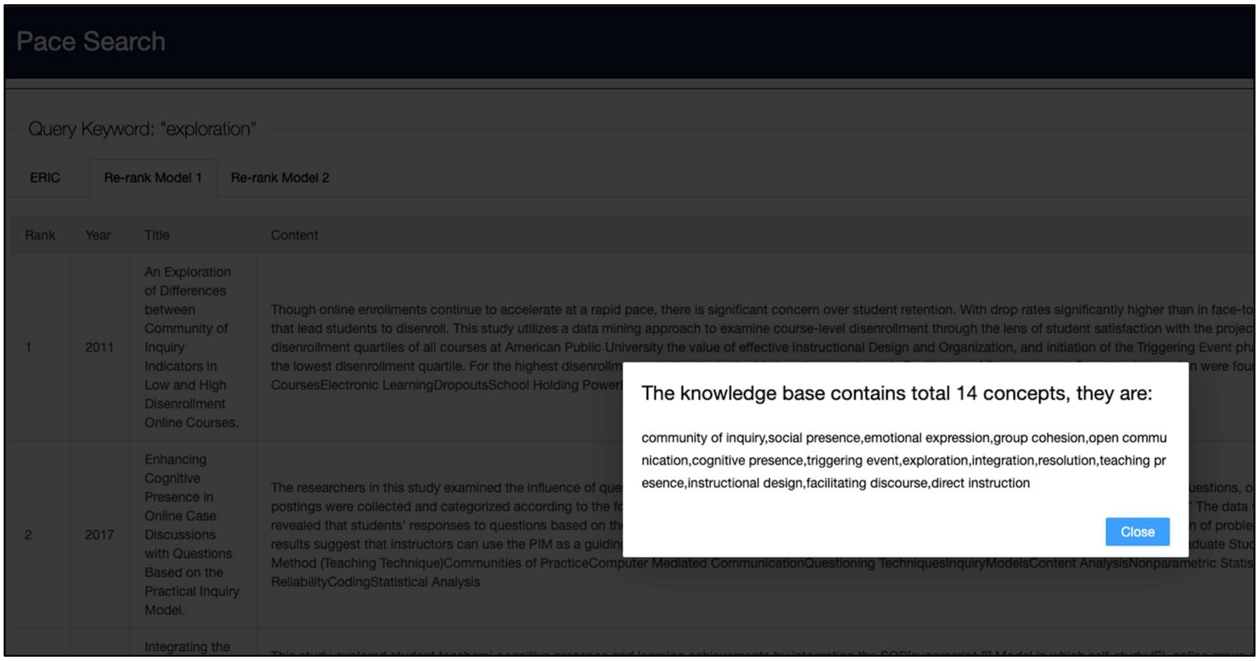
Query Keyword: "exploration"

Rank	Year	Title	Content
1	2018	Contextualizing Technology in the Classroom via Remote Access: Using Space Exploration Themes and Scanning Electron Microscopy as Tools to Promote Engagement in Geology/Chemistry Experiments.	A multidisciplinary science experiment was performed in K-12 classrooms focusing on the interconnection between technology with geology and chemistry. The engagement and passion for science of over eight hundred students across twenty-one classrooms, utilizing a combination of hands-on activities using relationships between Earth and space rock studies, followed by a remote access session wherein students remotely employed the use of a scanning electron microscope (SEM) and energy-dispersive spectroscopy (EDS) to validate their findings was investigated. Participants represent predominantly low-income minority communities, with little exposure to the themes and equipment used, despite being freely available resources. Students indicated greatly increased interest in scientific practices and careers, as well as a better grasp of the content as a result of the lab and remote access coupling format.Science InstructionLaboratory EquipmentScience ExperimentsGeologyChemistryTechnology Uses in EducationHands on SciencePretests PosttestsSpectroscopyElementary School StudentsMiddle School StudentsHigh School StudentsSpace ExplorationElementary School TeachersMiddle School TeachersSecondary School TeachersStudent SurveysCalifornia (Los Angeles)
2	2019	Investigation of the Knowledge and Views of Seventh Graders on the Relationship between Space Research and Technological Developments.	The main purpose of this study is to examine the views of seventh grade students on the relationship between space research and technological developments. The mixed method was used in the research. The true-false test was used to collect quantitative data. Qualitative data were collected through a fully structured interview form. Homogeneous samples were used in the study involving 134 participants. The research was conducted with 7th grade students in a public school in the Eastern Anatolia Region during the 2018-2019 academic years. Data were analyzed with the help of Microsoft Excel programs. Descriptive and content analysis and a technique such as frequency and percentage were used. With the interpretation of the findings, it was found that seventh grade students had very good knowledge about space research and technological developments. Students argue that space research has an impact on technological developments; technological research also has an impact on space exploration. Suggestions are presented in line with the results obtained.Grade 7Foreign CountriesSpace SciencesTechnological AdvancementSpace ExplorationKnowledge LevelStudent AttitudesTurkey
3	2019	Revisiting Second Graders' Robotics with an Understand/Use-Modify-Create (U[superscript 2]MC) Strategy.	This study, a sub-study of a National Science Foundation (NSF) funded research project, applies a modified strategy of the U[superscript 2]MC for an eight-week afterschool robotics curriculum to promote upper elementary students' computational thinking in the second grade. Twenty-one students in second grade participated in a Life on Mars project which lasted for ten days with one class hour per day. They participated in activities learning coding concepts, basics of robotics, as well as exploring life on Mars. Most notably, the study found a significant increase in participants' computational thinking skills. In addition, participants came to understand basic robotics, including operation, composites, and codes. Implications for future research and robotics curriculum design are discussed in the presentation.Grade 2RoboticsElementary School StudentsElementary School ScienceThinking SkillsMental ComputationProblem SolvingCodingProgrammingHands on ScienceSTEM EducationTeaching MethodsCurriculum DesignObject ManipulationScience ActivitiesScience EducationSpace ExplorationIdaho (Boise)

Figure 49 The prototype's interface

From the above figure, we can observe that three tabs are used to select and display the search results. Among them, re-rank model 1 and re-rank model 2 refer to the different retrieval results returned by the proposed methods based on the different knowledge bases. When the user selects a specific tab, the search results returned by the corresponding method will be displayed in a list.

At the same time, the prototype interface was designed to display the summary information of the introduced database. When the user selects re-rank model 1 or re-rank model 2, the interface will pop up the summary information of the knowledge base introduced by the model, including the number of the concepts in the knowledge base and the term name of these concepts (the design is shown in Figure 50 and 51).



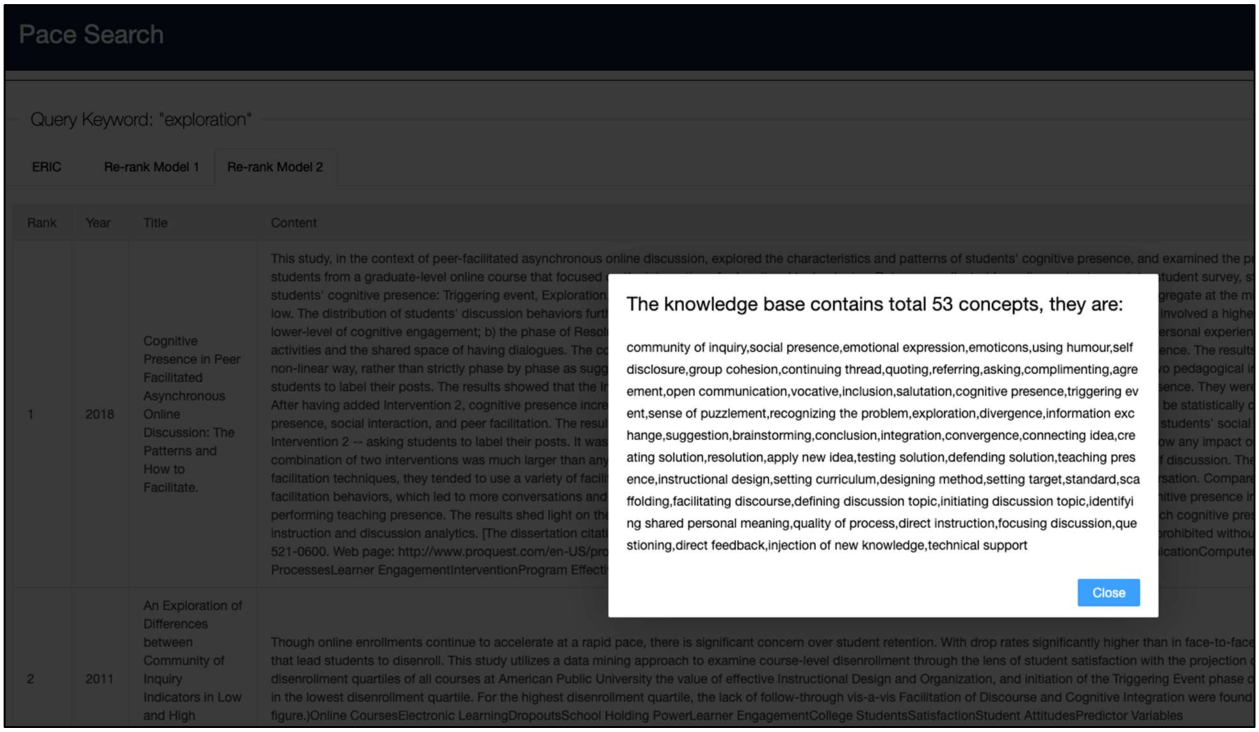
Query Keyword: "exploration"

Rank	Year	Title	Content
1	2011	An Exploration of Differences between Community of Inquiry Indicators in Low and High Disenrollment Online Courses.	Though online enrollments continue to accelerate at a rapid pace, there is significant concern over student retention. With drop rates significantly higher than in face-to that lead students to disenroll. This study utilizes a data mining approach to examine course-level disenrollment through the lens of student satisfaction with the projec disenrollment quartiles of all courses at American Public University the value of effective Instructional Design and Organization, and initiation of the Triggering Event pha the lowest disenrollment quartile. For the highest disenrollm n were fou CoursesElectronic LearningDropoutsSchool Holding Powerl
2	2017	Enhancing Cognitive Presence in Online Case Discussions with Questions Based on the Practical Inquiry Model.	The researchers in this study examined the influence of que uestions, o postings were collected and categorized according to the fo ' The data revealed that students' responses to questions based on th n of proble results suggest that instructors can use the PIM as a guidin duate Stuc Method (Teaching Technique)Communities of PracticeComputer Mediated CommunicationQuestioning TechniquesInquiryModelsContent AnalysisNonparametric Statis ReliabilityCodingStatistical Analysis
		Integrating the	

The knowledge base contains total 14 concepts, they are:

community of inquiry,social presence,emotional expression,group cohesion,open communication,cognitive presence,triggering event,exploration,integration,resolution,teaching presence,instructional design,facilitating discourse,direct instruction

Figure 50 The summary information of knowledge base in re-rank mode 1



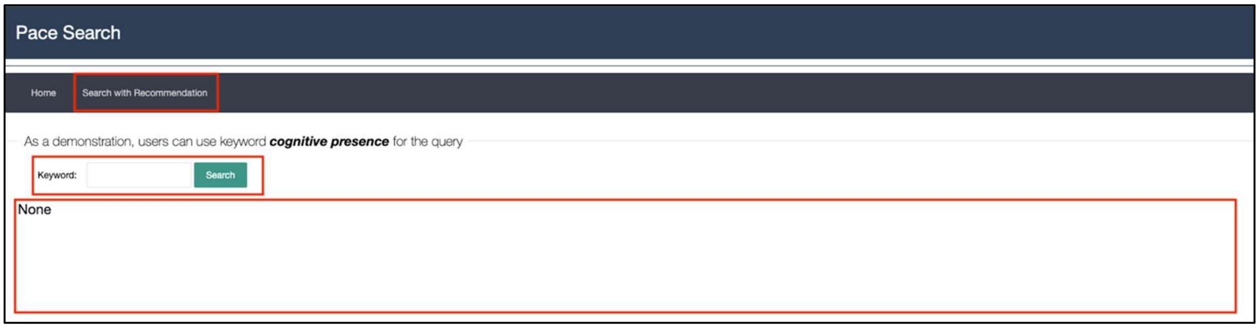
Rank	Year	Title	Content
1	2018	Cognitive Presence in Peer Facilitated Asynchronous Online Discussion: The Patterns and How to Facilitate.	This study, in the context of peer-facilitated asynchronous online discussion, explored the characteristics and patterns of students' cognitive presence, and examined the p students from a graduate-level online course that focused tudent survey, s students' cognitive presence: Triggering event, Exploration gregate at the m low. The distribution of students' discussion behaviors furt involved a highe lower-level of cognitive engagement; b) the phase of Resol ersonal experien activities and the shared space of having dialogues. The co ence. The results non-linear way, rather than strictly phase by phase as sugg o pedagogical i students to label their posts. The results showed that the In ence. They were After having added Intervention 2, cognitive presence incre be statistically c presence, social interaction, and peer facilitation. The resul students' social Intervention 2 -- asking students to label their posts. It was ow any impact o combination of two interventions was much larger than any f discussion. The facilitation techniques, they tended to use a variety of facili rsation. Compare facilitation behaviors, which led to more conversations and nitive presence i performing teaching presence. The results shed light on the ch cognitive pres instruction and discussion analytics. [The dissertation citat prohibited withou 521-0600. Web page: http://www.proquest.com/en-US/pro icationComputer ProcessesLearner EngagementInterventionProgram Effecti
2	2011	An Exploration of Differences between Community of Inquiry Indicators in Low and High	Though online enrollments continue to accelerate at a rapid pace, there is significant concern over student retention. With drop rates significantly higher than in face-to-face that lead students to disenroll. This study utilizes a data mining approach to examine course-level disenrollment through the lens of student satisfaction with the projection o disenrollment quartiles of all courses at American Public University the value of effective Instructional Design and Organization, and initiation of the Triggering Event phase o in the lowest disenrollment quartile. For the highest disenrollment quartile, the lack of follow-through vis-a-vis Facilitation of Discourse and Cognitive Integration were found figure.)Online CoursesElectronic LearningDropoutsSchool Holding PowerLearner EngagementCollege StudentsSatisfactionStudent AttitudesPredictor Variables

The knowledge base contains total 53 concepts, they are:

community of inquiry,social presence,emotional expression,emoticons,using humour,self disclosure,group cohesion,continuing thread,quoting,referring,asking,complimenting,agreement,open communication,vocative,inclusion,salutation,cognitive presence,triggering event,sense of puzzlement,recognizing the problem,exploration,divergence,information exchange,suggestion,brainstorming,conclusion,integration,convergence,connecting idea,creating solution,resolution,apply new idea,testing solution,defending solution,teaching presence,instructional design,setting curriculum,designing method,setting target,standard,scaffolding,facilitating discourse,defining discussion topic,initiating discussion topic,identifying shared personal meaning,quality of process,direct instruction,focusing discussion,questioning,direct feedback,injection of new knowledge,technical support

Figure 51 The summary information of knowledge base in re-rank mode 2

On the search page with a recommendation system, the user can search the information by entering the query keyword in the text input box. When the user finished entering the keywords, the search results are displayed at the bottom of the interface. As a demonstration, users can use cognitive presence as an example to complete the test. Figure 52 is the design of the search page with the recommendation system.

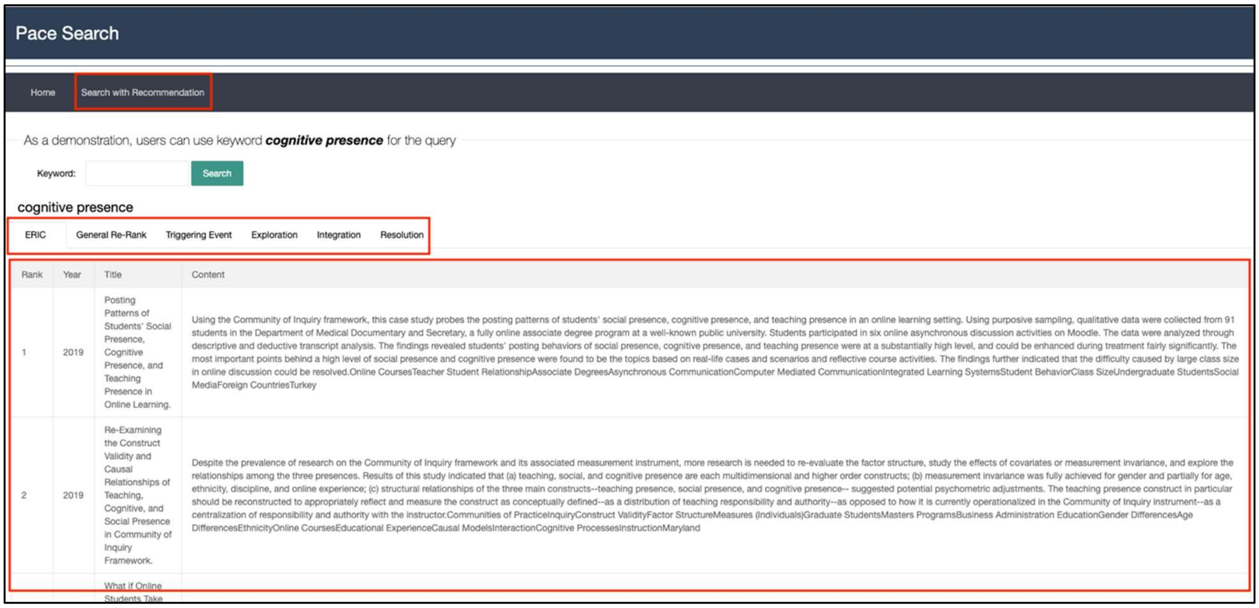


As a demonstration, users can use keyword ***cognitive presence*** for the query

None

Figure 52 The design of the search page with recommendation system

After the user finished entering the keywords, the search page returns several tabs according to the recommendation system. These tabs correspond to the some more accurate search range recommended by the knowledge base. For example, through the knowledge base established for the experiment, we can know that triggering events, exploration, integration, and resolution are all cases of cognitive presence. Therefore, the search page with the recommendation system will generate the following tabs: ERIC, General Re-rank, Triggering event, Exploration, Integration, and Resolution. In the ERIC tab, the page will return the query results from the ERIC. In the General Re-Rank tab, the re-ranked query result by the re-rank method proposed will be returned. The remaining tabs will return search results with a more precise search range generated by the recommendation system. Users can select the corresponding tab according to their needs to browse the search results as they wish. Figure 53 is the search result on the search page with the recommendation system.



As a demonstration, users can use keyword ***cognitive presence*** for the query

Keyword:

cognitive presence

Rank	Year	Title	Content
1	2019	Posting Patterns of Students' Social Presence, Cognitive Presence, and Teaching Presence in Online Learning.	Using the Community of Inquiry framework, this case study probes the posting patterns of students' social presence, cognitive presence, and teaching presence in an online learning setting. Using purposive sampling, qualitative data were collected from 91 students in the Department of Medical Documentary and Secretary, a fully online associate degree program at a well-known public university. Students participated in six online asynchronous discussion activities on Moodle. The data were analyzed through descriptive and deductive transcript analysis. The findings revealed students' posting behaviors of social presence, cognitive presence, and teaching presence were at a substantially high level, and could be enhanced during treatment fairly significantly. The most important points behind a high level of social presence and cognitive presence were found to be the topics based on real-life cases and scenarios and reflective course activities. The findings further indicated that the difficulty caused by large class size in online discussion could be resolved.Online CoursesTeacher Student RelationshipAssociate DegreesAsynchronous CommunicationComputer Mediated CommunicationIntegrated Learning SystemsStudent BehaviorClass SizeUndergraduate StudentsSocial MediaForeign CountriesTurkey
2	2019	Re-Examining the Construct Validity and Causal Relationships of Teaching, Cognitive, and Social Presence in Community of Inquiry Framework.	Despite the prevalence of research on the Community of Inquiry framework and its associated measurement instrument, more research is needed to re-evaluate the factor structure, study the effects of covariates or measurement invariance, and explore the relationships among the three presences. Results of this study indicated that (a) teaching, social, and cognitive presence are each multidimensional and higher order constructs; (b) measurement invariance was fully achieved for gender and partially for age, ethnicity, discipline, and online experience; (c) structural relationships of the three main constructs--teaching presence, social presence, and cognitive presence-- suggested potential psychometric adjustments. The teaching presence construct in particular should be reconstructed to appropriately reflect and measure the construct as conceptually defined--as a distribution of teaching responsibility and authority--as opposed to how it is currently operationalized in the Community of Inquiry instrument--as a centralization of responsibility and authority with the instructor.Communities of PracticeInquiryConstruct ValidityFactor StructureMeasures (Individuals)Graduate StudentsMasters ProgramsBusiness Administration EducationGender DifferencesAge DifferencesEthnicityOnline CoursesEducational ExperienceCausal ModelsInteractionCognitive ProcessesInstructionMaryland
		What if Online Students Take	

Figure 53 The search results in the page with recommendation system

4.4 Query Testing

Compared the results obtained by different re-rank methods with the search results returned by ERIC, we can intuitively observe the retrieval capabilities of different methods. Based on the experiment's assumption, the proposed improvement method supported by domain knowledge should have better retrieval ability than keyword-based searches in a particular field. More specifically, the search engine is expected to return the documents that contain more concepts related to the query keyword via the ontology-based search and knowledge-graph based search, and then rank those documents in the top. For example, in the demonstration, we can observe that some documents are highly semantic related to the query but low rank in the search result from the ERIC. However, when the knowledge base was introduced in the search engine (ontology-based model or knowledge graph-based model), this problem has been improved significantly. Table 2 shows the first document in different results (ERIC, Re-rank method 1, and Re-rank method 2). The concepts related to the query (exploration) from the knowledge base are bolded. From the results, we can see that with the introduction of the knowledge base, the search engine not only retrieves the query when retrieving related documents but also retrieves semantics concepts related to query. For example, in the first article among the search results returned by ERIC, we can observe that it only retrieved the keyword "exploration." However, the re-ranked methods can retrieve the semantically related concepts with the query, such as “community inquiry,” “cognitive presence.” From the experiment's results, we can understand that as knowledge expression becomes more and more powerful, the ability of search engines to identify and retrieve semantic-related concepts is growing. The experiment demonstrated the strong potential of the search engine on the identification of semantic similar concepts expressions when the domain-specific knowledge representation was introduced.

Table 2 Examples of matched documents by different models. Bold text is the related concept.

Query	Methods	Publish Year	Document
exploration	ERIC	2018	Title: Contextualizing Technology in the Classroom via Remote Access: Using Space **Exploration** Themes and Scanning Electron Microscopy as Tools to Promote Engagement in Geology/Chemistry Experiments. Abstract: A multidisciplinary science experiment … access coupling format. Keywords: Science Instruction, …, Space **Exploration**, …
	Re-rank 1	2011	Title: An **Exploration** of Differences between **Community of Inquiry** Indicators in Low and High Disenrollment Online Courses.

			Abstract: Though online enrollments, … with the projection of Teaching, Social and **Cognitive Presence**. … initiation of the **Triggering Event** phase of **Cognitive Presence** …
	Re-rank 2	2018	Title: **Cognitive Presence** in Peer Facilitated Asynchronous Online Discussion: The Patterns and How to Facilitate. Abstract: This study, in the context of peer-facilitated asynchronous online discussion, explored the characteristics and patterns of students' **cognitive presence**, and examined the practices that aim to enhance **cognitive presence** development. … Results demonstrated four phases of students' **cognitive presence**: **Triggering event**, **Exploration**, **Integration**, and **Resolution**. Among the four phases, students' **cognitive presence** tended to aggregate at the middle phases: **Integration** and **Exploration**. Percentage of the **Resolution** was very low. … **Integration** and **Resolution** involved a higher-level of cognitive engagement, and **Triggering event** and **Exploration** involved a

			lower level … 1) providing guidance on peer facilitation techniques; 2) **asking** students to label their posts … especially when students are coached in performing **teaching presence.**

4.5 Experiment Evaluation

Based on the experimental assumptions, the proposed methods should be able to:

- Rank the semantically related document to the top.
- Collect the semantically related document even they do not contain the query keyword.

Table 2 lists the first result from different methods, respectively. From this table, we can observe that with the knowledge base's expansion, the articles which are ranked in the top include more concepts in the knowledge base. Moreover, the concepts are more semantically, similar to the query keyword. The articles which are semantically related to query keyword can be minded out and ranked to the top by using proposed methods in the dissertation.

From a professional perspective, experts in education have been involved in the experiment. They compared the performance between the existing search engine and the re-ranking models. Because this experiment is involving human participants, so that I have completed the CITI Program (Collaborative Institutional Training Initiative)'s training and obtained the certificate. The Collaborative Institutional Training Initiative (CITI Program) is dedicated to promoting the public's trust in the research enterprise by providing high quality, peer-reviewed, web-based educational courses in research, ethics, regulatory oversight, responsible conduct of research, research administration, and other topics pertinent to the interests of member organizations and individual learners. These materials are designed and regularly updated to:

- Enhance the knowledge and professionalism of investigators, staff, and students researching in the United States and internationally
- Educate members, administrators, and leadership of ethics committees that review and oversee research
- Promote ethical research at organizations through the education of research administrators and organizational leadership

The volunteers have completed the designed questionnaire. And the questionnaire contains the following sections:

- Different return lists from existing search engine platforms and reranking models, and the lists are not identified for participants.
- The participants will review different lists and evaluate the lists based on the survey questions.

The experiment used "exploration," which is one of the concepts in "cognitive presence" as a keyword to retrieve academic articles in ERIC. The questionnaire consists of three lists. ERIC and the two proposed methods based on different knowledge bases generated their orders, respectively. Rating the list by using 1 – 6, 1 refers to the poor, and 6 refers to the excellent. Moreover, the survey questions were designed for making a comparison among the three lists in the following:

1. To what extent, the list covers the relevant information with the search inquiry “Exploration.”
2. To what extent, the list contains the highest proportion of high-quality papers.
3. To what extent, the list contains the most relevant paper about “Exploration” among the three lists.
4. To what extent, the list would effectively help to target the search inquiry “Exploration” in the education domain.
5. Overview, please rate the three lists of retrieval ranking.

4.6 Questionnaire results

Five volunteers from the School of Education in a top private university have participated in this research, and they have composed of three Ph.D. candidates and two master students. Retrieval ranking list ERIC generates one, retrieval ranking list two is generated by Re-rank model 1, and retrieval ranking list three is generated by Re-rank model 2. The difference between Re-rank model 1 and Re-rank model 2 is that they use different knowledge bases, respectively. Moreover, the knowledge base used in Re-rank model 2 contains more concepts.

For the first question in the questionnaire, all the 5 participants agreed that the articles in list 3 contained the most relevant information with the query keyword “exploration.”

For the second question, four of the five participants considered the articles in the third list to be of high quality.

For the third question, all 5 participants considered that the third list contains the most relevant paper about “Exploration” among the three lists.

All the 5 participants rated that list three would effectively help them to target the search inquiry “Exploration” in the education domain. The evaluation scores given by the five participants are shown in the table below.

Table 3 The results review for the questionnaire

	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Question 1	List1: 2	List1: 2	List1: 2	List1: 1	List1: 1
	List2: 4	List2: 5	List2: 4	List2: 4	List2: 4
	List3: 6	List3: 6	List3: 5	List3: 6	List3: 5
Question 2	List1: 2	List1: 3	List1: 5	List1: 2	List1: 3
	List2: 4	List2: 3	List2: 4	List2: 4	List2: 4
	List3: 5	List3: 4	List3: 4	List3: 6	List3: 4
Question 3	List1: 2	List1: 3	List1: 3	List1: 2	List1: 2
	List2: 4	List2: 4	List2: 5	List2: 5	List2: 4

	List3: 4	List3: 5	List3: 6	List3: 5	List3: 6
Question 4	List1: 2	List1: 3	List1: 3	List1: 4	List1: 4
	List2: 5	List2: 4	List2: 5	List2: 4	List2: 5
	List3: 5	List3: 6	List3: 6	List3: 5	List3: 6
Question 5	List1: 2	List1: 2	List1: 2	List1: 4	List1: 4
	List2: 4	List2: 4	List2: 5	List2: 5	List2: 5
	List3: 5	List3: 6	List3: 6	List3: 6	List3: 6

The evaluation results of all participants are calculated and obtained the following results:

1. Question 1: The average score of List 1 is 1.6, the average score of List 2 is 4.2, and the average score of List 3 is: 5.6.
2. Question 2: The average score of List 1 is 3, the average score of List 2 is 3.8, and the average score of List 3 is: 4.6.
3. Question 3: The average score of List 1 is 2.4, the average score of List 2 is 4.4, and the average score of List 3 is: 5.2.
4. Question 4: The average score of List 1 is 3.2, the average score of List 2 is 4.6, and the average score of List 3 is: 5.6.
5. Question 5: The average score of List 1 is 2.8, the average score of List 2 is 4.6, and the average score of List 3 is: 5.8.

By comparing the evaluations of the participants, we can find that List 3 received the highest score in each question, and List 2 received the second-highest score. In other words, the search results re-ranked by the method proposed in this dissertation can be recognized by users. As knowledge expression becomes more and more powerful, the ability of search engines to identify and retrieve semantic-related concepts is growing, and the performance of the search engine was improved. The experiment demonstrates the strong potential of the search engine on the identification of semantic similar concepts expressions when the domain-specific knowledge representation was introduced.

From the survey results, it confirmed that the experimental assumptions are valid and proved that the methods proposed in this dissertation could improve the current traditional keyword-based search and emulate the effects of semantic search. The proposed methods can improve the lack of ability of the traditional search engine in semantic understanding and avoid inefficiencies.

Chapter 5 Conclusion

Although the semantic search engine can execute the extensive and effective semantic reasoning in the network environment or documents by annotating the resource objects, annotating every web page or document is extremely time-consuming for the search engine. This dissertation designed a multi-promoting approach to improve the current traditional keyword-based search and emulate the effects of semantic search. The ontology and knowledge graph are described as domain knowledge, respectively. Then, the search engine can understand users more semantically when it obtains knowledge.

Two challenges are solved in this dissertation. One is to convert knowledge into a machine-readable quantification result. Then, the search engines could apply the domain knowledge in continuous spaces. Another one is to let the search engines to learn the relationship between concepts from the domain knowledge. When the users try to search one of the concepts in the domain knowledge, the search engine can return the result not only the corresponding queried concept, but also its relevant concepts according to the degree of association. In this approach, we can improve the search engine the capabilities to capture the conceptualizations involved in users' intention and web pages' content meaning. Because the semantic relevance of the keywords and another vocabulary in the web pages are considered, it can optimize the user experience and improve the performance of domain knowledge in search engines. This dissertation designed a recommendation system that can help users narrow down the search scope during the retrieval process based on the introduced domain knowledge.

Through the designed experiment, the proposed methods are expected to achieve two desired outcomes. First, the re-ranking algorithm should be able to rank the semantically related document to the top. Second, the re-ranking algorithm should be able to collect the semantically related document even they don’t contain the query keyword. Based on these two assumptions, a prototype was designed and conducted for the test. The experiment used academic articles from ERIC as the experimental data set. It used “exploration,” which is one of the concepts in “cognitive presence” as a keyword to retrieve academic articles in ERIC. From the experimental questionnaire conducted by experts, the results confirmed that the experimental hypothesis is correct, and they can improve the lack of ability of the traditional search engine in semantic understanding and avoid inefficiencies.

With the deepening of the research, the method proposed in this dissertation also has potential limitations. In this dissertation, two proposed approaches could convert the domain knowledge into a machine-readable quantification result. Then, the search engines can learn the relationship between concepts from the domain knowledge. However, the proposed methods may not suit every situation. The domain knowledge for the proposed approaches depends on well-developed expertise in a specific domain. The experts in this domain established the domain knowledge bases based on their experience and professional academic subject development. Some non-academic information could not be built into a knowledge base, such as the business data on the Internet.

On the other hand, the proposed re-ranking algorithms' time complexity determined by the capacity of the database and the size of the candidate documents in the database. The algorithms should check every document in the database and calculate the semantic similarity between the query keyword and the content in the document. Thus, the proposed approaches in this dissertation are more suitable for the search in a specific domain than a general search, such as academic retrieval.

In the future, I am going to apply the neural network to the introduced domain knowledge (e.g., Graph Neural Network) to optimize the embedded knowledge. Then the search engine could obtain more accuracy semantic similarity from the concepts among the domain knowledge. Also, neural networks can be used to predict the entities and relationships among the domain knowledge, thereby optimizing the recommendation system in the query and alleviating the potential limitation in the proposed approaches.

References

[1] L. P. Sergey Brin, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, vol. 30, no. 1-7, pp. 107-117, 1998.

[2] J. M. Kleinberg, "Hubs, authorities, and communities," ACM Computing Surveys, vol. 31, no. 4, 1999.

[3] T. J. H. O. L. Berners-Lee, "The semantic web," Scientific american, vol. 284, no. 5, pp. 28-37, 2001.

[4] C. T. H. T. B.-L. Bizer, "Linked data: The story so far," in Semantic services, interoperability and web applications: emerging concepts, IGI Global, 2011, pp. 205-227.

[5] T. Berners-Lee, "Linked data-design issues (2006)," 2011.

[6] L. T. B.-L. a. R. T. F. Masinter, "Uniform resource identifier (URI): Generic syntax," 2005.

[7] R. e. a. Fielding, "Hypertext transfer protocol–HTTP/1.1," 1999.

[8] H. e. a. Knublauch, "A semantic web primer for object-oriented software developers," W3c working group note, W3C, 2006.

[9] N. Luhmann, Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität., Stuttgart: Ferdinand Enke Verlag, 1989.

[10] H. A. Simon, "Rational choice and the structure of the environment," Psychological Review, vol. 63, no. 2, pp. 129-138, 1956.

[11] R. S. Gerd Gigerenzer, Bounded Rationality: The Adaptive Toolbox, Cambridge: MIT Press, 2002.

[12] B. P. T. J. Helene Hembrooke, "In Google We Trust: Users' Decisions on Rank, Position, and Relevance," Journal of Computer-Mediated Communication, vol. 12, no. 3, pp. 801-823, 2007.

[13] J. F. Allen, "Maintaining knowledge about temporal intervals," in Readings in qualitative reasoning about physical systems, Morgan Kaufmann, 1990, pp. 361-372.

[14] H. J. Levesque, "Knowledge representation and reasoning," Annual review of computer science, vol. 1, no. 1, pp. 255-287, 1986.

[15] M. M.-L. M. Chein, Graph-based knowledge representation: computational foundations of conceptual graphs, Springer Science & Business Media, 2008.

[16] N. T. B.-L. W. H. Shadbolt, "The semantic web revisited," IEEE intelligent systems, vol. 21, no. 3, pp. 96-101, 2006.

[17] S. R. S. J. J. M. S´ebastien Harispe, Semantic Similarity from Natural Language and Ontology analysis, Morgan & Claypool, 2015.

[18] M. Leong, "Measuring the semantic relatedness between words and images," in Proceedings of the 9th International Conference on Computational Semantics, 2011.

[19] H. Budanitsky, " Evaluating Wordnet-based measures of lexical semantic relatedness," Computational Linguistics, vol. 32, no. 1, pp. 13-47, 2006.

[20] G. M. E. L. Salton, "Computer evaluation of indexing and text processing," Journal of the ACM (JACM), vol. 15, no. 1, pp. 8-36, 1968.

[21] J. D. U. Anand Rajaraman, Mining of Massive Datasets, Cambridge University Press, 2011.

[22] B. G. S. L. C. B. Joeran Beel, "Research-paper recommender systems: a literature survey," International Journal on Digital Libraries, vol. 17, no. 4, pp. 305-338, 2016.

[23] L. Yu, Introduction to the Semantic Web and Semantic Web Services, Chapman and Hall/CRC, 2007.

[24] Hou Yu, T. L. "Semantic-Based Resume Screening System," in Advances in Intelligent Systems and Computing, vol. 880, Springer, Cham, 2018, pp. 649-658.

[25] Hou Yu, T. L. "An Ontology-based Ranking Model in Search Engines," Journal of Computer Science Research, vol. 1, no. 2, pp. 8-16, 2019.

[26] K. Bollacker, C. Evans, P. Paritosh, T. Sturge and and Taylor, "Freebase: A collaboratively created graph database," in Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, 2008.

[27] G. A. Miller, "Wordnet: A lexical database for english," Communications of the ACM, vol. 38, no. 11, pp. 39-41, 1995.

[28] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig and e. al, " Gene ontology: Tool for the unification of biology," Nature genetics, vol. 25, no. 1, pp. 25-29, 2000.

[29] N. U. A. G.-D. J. W. O. Y. Antoine Bordes, "Translating Embeddings for Modeling Multi-relational Data," in Advances in Neural Information Processing Systems 26, 2013.

[30] J. Z. J. F. Z. C. ZhenWang, "Knowledge Graph Embedding by Translating on Hyperplanes," in Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[31] D. R. T. A. a. W. A. Garrison, "Critical inquiry in a text-based environment: Computer conferencing in higher education."," The internet and higher education, vol. 2, no. 2-3, pp. 87-105, 1999.

[32] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition - Special issue: Current issues in knowledge modeling, vol. 5, no. 2, pp. 199-220, 1993.

[33] V. R. B. Asuncion Gomez Perez, "Overview of Knowledge Sharing and Reuse Components: Ontologies and Problem-Solving Methods," in Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5), Stockholm, 1999.

[34] I. D. L. T. N. J. Khagesh Patel, "Extending OWL to Support Custom Relations," in 2015 IEEE 2nd International Conference on Cyber Security and Cloud Computing, 2015.

[35] M. E. DeBakey, "The National Library of Medicine: Evolution of a Premier Information Center," JAMA Network, vol. 266, pp. 1252-1258, 1991.

[36] N. F. N. N. H. S. P. R. A. C. N. T. T. M. A. M. Patricia L. Whetzel, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications," Nucleic Acids Research, vol. 39, pp. W541-545, 2011.

[37] T.-J. a. S. L. M. a. C. Q.-R. a. C. M. a. C. D. J. a. F. R. a. H. Y. a. K. W. A. a. K. H. a. M. D. a. o. Wu, "Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis," Database (Oxford), vol. 2015, 2015.

[38] “Knowledge representation and reasoning -Wikipedia.” [Online]. Available: https://en.wikipedia.org/wiki/Knowledge_representation_and_reasoning.
[Accessed: April 12, 2017].

[39] Ontology-Based Integration of Information — A Survey of Existing Approaches. H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner

[40] T. R. Gruber. Toward principles for the design of ontologies used for knowledge haring. Presented at the Padua workshop on Formal Ontology, March 1993, later published in International Journal of Human-Computer Studies, Vol. 43, Issues 4-5, November 1995, pp. 907-928.

[41] P. Devanbu, R. J. Brachman, P.G. Selfridge, and B.W. Ballard. LASSIE: A knowledge-based software information system. Communications of the ACM, 34(5):34-49, 199

[42] C. Vergara-Niedermayr, F. Wang, T. Pan, T. Kurc, and J. Saltz, “SEMANTICALLY INTEROPERABLE XML DATA,” Int. J. Semantic Comput., vol. 07, no. 03, pp. 237–255, Sep. 2013.

[43] W. Van, N. A Haider, P. C. Roy, A. M. Ahmad, and S. S.R. Abidi. “A Comparison of Mobile Rule Engines for Reasoning on Semantic Web Based Health Data,” 126–33. IEEE, 2014. doi:10.1109/WI-IAT.2014.25.

[44] L. Tao, S. Golikov, K. Gai, M. Qiu “A Reusable Software Component for Integrated Syntax and Semantic Validation for Services Computing” The 9th International IEEE Symposium on Service- Oriented System Engineering, At San Francisco Bay, CA, USA

[45] L. Tao “Extending OWL with Custom Relations for Knowledge-Driven Intelligent Agents” 15th German Conference, MATES 2017, Leipzig, Germany, August 23–26, 2017

[46] M. Sette, L. Tao, N. Jiang “A Knowledge-Driven Web Tutoring System Framework for Adaptive and Assessment-Driven Open-Source Learning” The IEEE 3rd international conference on Cyber Security and Cloud Computing, At New York City, NY, USA

[47] J. Hu and L. Tao, “An Extensible Constraint Markup Language: Specification, Modeling, and Processing”. XML 2004 Proceedings by SchemaSoft, 2004

[48] J. Hu and L. Tao. "Visual modeling of XML constraints based on a new Extensible Constraint Markup Language," Engineering Letters, Issue v13-3, December 2006. pp.248-254.

[49] L. Tao and S. Golikov, “Integrated Syntax and Semantic Validation for Services Computing.” http://csis.pace.edu/~ctappert/srd2013/d2.pdf [Accessed: 29-July-2015].

[50] L. Tao, N. Jiang "A Practical Guide to Building OWL Ontologies and Knowledge Graphs Using Protégé 5 Extended by Pace University" [Accessed: 07-Apr-2016].

Appendix A Retrieval Ranking Questionnaire

The major of participant: ____

The education level of participant: _______________________________________

Date: ____

This dissertation designed a multi-promoting approach to improve the current traditional keyword-based search and emulate the effects of semantic search. The proposed methods are expected to improve the lack ability of the traditional search engine in semantic understanding and avoid inefficiencies. The experimental data set consists of academic articles from ERIC. The experiment used ‘exploration’ which is one of the concepts in ‘cognitive presence’ as a keyword to retrieve academic articles in ERIC. This survey consists of three lists, and their order was generated by ERIC and the methods which are proposed in this dissertation based on different knowledge bases.

After reviewing the three lists of retrieval ranking, please make a comparison among the three lists by answering the following questions. Rating the list by using 1 – 6, 1 refers to the poor and 6 refers to the excellent.

1. To what extent, the list covers the relevant information with the search inquiry “Exploration”.
2. To what extent, the list contains the highest proportion of high-quality papers.

3. To what extent, the list contains the most relevant paper about "Exploration" among the three lists.
4. To what extent, the list would effectively help you to target the search inquiry "Exploration" in education domain.
5. Overview, please rate the three lists of retrieval ranking.

Retrieval Ranking List 1:

Rank	Year	Title	Source	Type	Content
1	2018	Contextualizing Technology in the Classroom via Remote Access: Using Space Exploration Themes and Scanning Electron Microscopy as Tools to Promote Engagement in	Journal of Technology and Science Education.	Journal Articles.	A multidisciplinary science experiment was performed in K-12 classrooms focusing on the interconnection between technology with geology and

Rank	Year	Title	Source	Type	Content
		Geology/Chemistry Experiments.			chemistry. The engagement and passion for science of over eight hundred students across twenty-one classrooms, utilizing a combination of hands-on activities using relationships between Earth and space rock studies, followed by a remote access session wherein students remotely employed the use of a scanning

Rank	Year	Title	Source	Type	Content
					electron microscope (SEM) and energy-dispersive spectroscopy (EDS) to validate their findings was investigated. Participants represent predominantly low-income minority communities, with little exposure to the themes and equipment used, despite being freely available

Rank	Year	Title	Source	Type	Content
					resources. Students indicated greatly increased interest in scientific practices and careers, as well as a better grasp of the content as a result of the lab and remote access coupling format. Keywords: Science Instruction Laboratory Equipment

Rank	Year	Title	Source	Type	Content
					Science Experiments Geology Chemistry Technology Uses in Education Hands on Science Pretests Posttests Spectroscopy Elementary School Students Middle School Students High School Students Space Exploration

Rank	Year	Title	Source	Type	Content
					Elementary School Teachers Middle School Teachers Secondary School Teachers Student Surveys
2	2019	Investigation of the Knowledge and Views of Seventh Graders on the Relationship between Space Research and Technological Developments.	Universal Journal of Educational Research.	Journal Articles.	The main purpose of this study is to examine the views of seventh grade students on the relationship between space research and technological developments. The

Rank	Year	Title	Source	Type	Content
					mixed method was used in the research. The true-false test was used to collect quantitative data. Qualitative data were collected through a fully structured interview form. Homogeneous samples were used in the study involving 134 participants. The research was conducted with 7th grade students in a

Rank	Year	Title	Source	Type	Content
					public school in the Eastern Anatolia Region during the 2018-2019 academic years. Data were analyzed with the help of Microsoft Excel programs. Descriptive and content analysis and a technique such as frequency and percentage were used. With the interpretation of the findings, it was found that seventh grade

Rank	Year	Title	Source	Type	Content
					students had very good knowledge about space research and technological developments. Students argue that space research has an impact on technological developments; technological research also has an impact on space exploration. Suggestions are presented in line with the results obtained.

Rank	Year	Title	Source	Type	Content
					Keywords: Grade 7 Foreign Countries Space Sciences Technological Advancement Space Exploration Knowledge Level Student Attitudes
3	2019	Revisiting Second Graders' Robotics with an Understand/Use-Modify-Create (U[superscript 2]MC) Strategy.	European Journal of STEM Education.	Journal Articles.	This study, a sub-study of a National Science Foundation (NSF) funded research project, applies a modified strategy

Rank	Year	Title	Source	Type	Content
					of the U[superscript 2]MC for an eight-week afterschool robotics curriculum to promote upper elementary students' computational thinking in the second grade. Twenty-one students in second grade participated in a Life on Mars project which lasted for ten days with one class hour

Rank	Year	Title	Source	Type	Content
					per day. They participated in activities learning coding concepts, basics of robotics, as well as exploring life on Mars. Most notably, the study found a significant increase in participants' computational thinking skills. In addition, participants came to understand basic robotics, including operation,

Rank	Year	Title	Source	Type	Content
					composites, and codes. Implications for future research and robotics curriculum design are discussed in the presentation. Keywords: Grade 2 Robotics Elementary School Students Elementary School Science Thinking Skills

Rank	Year	Title	Source	Type	Content
					Mental Computation Problem Solving Coding Programming Hands on Science STEM Education Teaching Methods Curriculum Design Object Manipulation Science Activities Science Education Space Exploration

Rank	Year	Title	Source	Type	Content
4	2018	Life Beyond--A Program to Use Astrobiology to Teach Science and Advance Space Exploration through Prisons.	Journal of Correctional Education.	Journal Articles.	The field of astrobiology is concerned with the origin, evolution, and distribution of life in the Universe. It contains within it civilization-level questions such as: What is the future of humanity on Earth and can we successfully explore and settle other planets? As such, it offers an educational framework for

Rank	Year	Title	Source	Type	Content
					both teaching basic science and for engaging individuals in questions about how society can take on its biggest challenges and opportunities. Life Beyond is a collaboration between the UK Centre for Astrobiology and the Scottish Prison Service (SPS) to take astrobiology into the prison environment.

Rank	Year	Title	Source	Type	Content
					Using a pilot program across four Scottish prisons, a 4-week astrobiology course focused on designing a station for Mars was developed. Learning outcomes ranged from improvements in literacy, numeracy, and science skills to enhancing civic responsibilities. The results of the

Rank	Year	Title	Source	Type	Content
					initiative are products such as Mars station designs, essays, and art, providing participants with tangible outputs. We describe the pilot initiative, the 4-week Life Beyond course, and draw conclusions about the use of astrobiology as a vehicle for teaching science and advancing social reform in

Rank	Year	Title	Source	Type	Content
					the prison environment. Keyword: Correctional Education Science Education Astronomy Biology Space Exploration Partnerships in Education Correctional Institutions Pilot Projects Foreign Countries

Rank	Year	Title	Source	Type	Content
					Outcomes of Education Institutionalized Persons Course Objectives Course Descriptions
5	2019	The Legacies of Apollo 11.	Physics Teacher.	Journal Articles.	Fifty years ago this summer, three men aboard Apollo 11 traveled from our planet to the Moon. On July 20, 1969, at 10:56:15 p.m. EDT, 38-year-old commander Neil

Rank	Year	Title	Source	Type	Content
					Armstrong moved his left foot from the landing pad of the lunar module (LM) Eagle onto the gray, powdery surface of the Sea of Tranquility and became the first person to step onto the lunar soil. Armstrong declared: "That's one small step for [a] man, one giant leap for mankind." Nineteen minutes later, 39-year-old LM pilot Edwin

Rank	Year	Title	Source	Type	Content
					"Buzz" Aldrin followed Armstrong onto the surface. Fifteen hours later, after spending two and a half hours outside of Eagle, the two men lifted off and returned to their command module (CM) Columbia, manned patiently by the third member of their crew, 38-year-old CM pilot Michael Collins. Four days later, the three men

Rank	Year	Title	Source	Type	Content
					were back home. Although five additional lunar landings would occur, each more challenging and scientifically ambitious than its predecessor, Apollo 11 stands alone as the greatest technological accomplishment of the 20th century. The mission also signaled the beginning of the end of the "Golden

Rank	Year	Title	Source	Type	Content
					Age" of America's space program. Keyword: Space Exploration Space Sciences
6	2019	The Apollo 1 Fire: A Case Study in the Flammability of Fabrics.	Physics Teacher.	Journal Articles.	This January marked the 52nd anniversary of the Apollo 1 fire. On Jan. 27, 1967, the interior of NASA's AS-204 command module (CM), occupied by American astronauts Roger Chaffee, Virgil

Rank	Year	Title	Source	Type	Content
					"Gus" Grissom, and Ed White, caught fire during a rehearsal of its scheduled Feb. 21 launch. By the time the ground crew was able to open the hatch, the three astronauts had perished. On April 24, 1967, NASA announced that the flight would be officially re-designated "Apollo 1." In this case study, we conduct a basic

Rank	Year	Title	Source	Type	Content
					horizontal flame test, patterned after the protocols set forth by the Environmental Protection Agency (EPA) to measure the ignitability of solids. The laboratory activity is a complementary exercise to the vertical flame test described in our previous article that examined the initial source of fuel for the fire that

Rank	Year	Title	Source	Type	Content
					destroyed the massive German zeppelin Hindenburg in 1937. Combining techniques from both case studies gives students a quantitative understanding of how the flammability of materials is tested and how a forensics approach to physics can be used to understand significant historical events.

Rank	Year	Title	Source	Type	Content
					Keyword: Science Instruction Physics Science Laboratories Space Exploration Scientific Concepts
7	2019	Using Real Data from the Kepler Mission to Find Potentially Habitable Planets: An Introductory	Physics Teacher.	Journal Articles.	A primary goal of general education introductory astronomy courses often is to provide students with examples of how

Rank	Year	Title	Source	Type	Content
		Astronomy Exercise.			science is actually done. Low to nonexistent mathematical prerequisites in some courses can make useful exercises difficult to find, and sometimes very difficult for students, especially if the exercises feature quantitative components. What follows describes an attempt to meet this goal through

Rank	Year	Title	Source	Type	Content
					the use of actual exoplanet data, available online, from the NASA Kepler Mission. The exercise described guides the students through an aspect of scientific investigation that they may otherwise not experience, the handling and analysis of a large set of actual scientific data. Keyword:

Rank	Year	Title	Source	Type	Content
					Astronomy Science Instruction Data Collection Space Exploration Space Sciences Inquiry Science Process Skills
8	2017	Have Astronauts Visited Neptune? Student Ideas about How Scientists Study the Solar System.	Journal of Astronomy & Earth Sciences Education.	Journal Articles.	The nature of students' ideas about the scientific practices used by astronomers when studying objects in our Solar System

Rank	Year	Title	Source	Type	Content
					is of widespread interest to discipline-based astronomy education researchers. A sample of middle-school, high-school, and college students (N = 42) in the U.S. were interviewed about how astronomers were able to learn about properties of the Solar System as a follow-up question after specific questions

Rank	Year	Title	Source	Type	Content
					about the nature of the Solar System and its objects. These students often held naive ideas about the practices of astronomy, and 19% of them proposed that humans or robots have returned samples of the planets to Earth for analysis. While the college students provided more sophisticated responses to the

Rank	Year	Title	Source	Type	Content
					questions than the younger students, we found that even they held naive ideas about human sample return and infrequently appealed to studying objects at a distance using telescopes. We propose that students are not receiving specific instruction that allows them to investigate the tools and practices of astronomy,

Rank	Year	Title	Source	Type	Content
					which leads them to rely on their prior knowledge about science practices in other disciplines (e.g., geoscience) when queried about how scientists study the Solar System. This result implies that instruction around the limits of human and robotic spaceflight is needed to allow students to have a more scientific understanding of

Rank	Year	Title	Source	Type	Content
					the practices of astronomy in studying the Solar System. Keyword: Astronomy Scientists Middle School Students High School Students College Students Interviews Student Attitudes Misconceptions

Rank	Year	Title	Source	Type	Content
					Scientific Research Observation Space Exploration Investigations Inferences Causal Models
9	2016	Astronauts in Outer Space Teaching Students Science: Comparing Chinese and American Implementations of Space-to-Earth	European Journal of Science and Mathematics Education.	Journal Articles.	The purpose of this study was to investigate differences between science lessons taught by Chinese astronauts in a space shuttle and those taught

Rank	Year	Title	Source	Type	Content
		Virtual Classrooms.			by American astronauts in a space shuttle, both of whom conducted experiments and demonstrations of science activities in a microgravity space environment. The study examined the instructional structure and science topics coverage, as well as the methods employed for helping students

Rank	Year	Title	Source	Type	Content
					conceptualize scientific laws via experimental demonstrations and activities. The analysis of the lessons sampled in this study revealed three predominant themes for how both the Chinese and the American astronauts conceptualized the science topics (i.e., Health and Life in Space, Work and Career in Space, and Exploration in

Rank	Year	Title	Source	Type	Content
					Space and Earth Science). The analysis also examined how the teacher-student interactions were structured. Research findings suggest that under the appropriate conditions informal science education can play a distinct role in providing students with experiences of: (a) experiments unavailable in classroom settings,

Rank	Year	Title	Source	Type	Content
					and (b) explanations of these experiments by field-based scientists conducting original research. Keyword: Virtual Classrooms Science Education Science Instruction Space Exploration Informal Education

Rank	Year	Title	Source	Type	Content
					Instructional Design Foreign Countries Teacher Student Relationship Science Experiments STEM Education Scientists Teaching Methods Science Teachers Comparative Analysis Scientific Concepts

Rank	Year	Title	Source	Type	Content
10	2016	Microgravity Playscapes: Play in Long-Term Space Missions.	American Journal of Play.	Journal Articles.	The authors examine the potential impact of play on astronauts adapting to the extreme conditions of space travel. They cite research showing that well-trained astronauts, though in general physically fit and emotionally stable, can suffer from--among other things--boredom and sensory deprivation in the confines of the

Rank	Year	Title	Source	Type	Content
					microgravity capsules of space flight. Astronauts on duty, the authors argue, are overscheduled, understimulated, isolated, and--importantly--play deprived. Introducing play into space flight routines, they contend, keeps astronauts saner, boosts their morale, and provides leisure-time pleasure.

Rank	Year	Title	Source	Type	Content
					They discuss the importance of play and its uses in Ackermann's Whole Child Development Guide, which, they argue, is also suitable for adult space travelers. And they provide guidelines for designing a playscape in microgravity that taps the unique, inherently playful qualities of

Rank	Year	Title	Source	Type	Content
					weightlessness itself. Keywords: Space Exploration Space Sciences Psychological Patterns Play Morale Leisure Time Adults Physics Scientific Concepts Science Activities

Rank	Year	Title	Source	Type	Content
					Hands on Science Creativity Visualization Simulation Well Being Navigation Social Development Recreational Activities

Retrieval Ranking List 2:

Rank	Year	Title	Source	Type	Content
1	2011	An Exploration of Differences between Community of Inquiry Indicators in Low and High Disenrollment Online Courses.	Journal of Asynchronous Learning Networks.	Journal Articles.	Though online enrollments continue to accelerate at a rapid pace, there is significant concern over student retention. With drop rates significantly higher than in face-to-face classes it is imperative that online providers develop an understanding of factors that lead students to disenroll. This study utilizes a data mining approach to examine course-level disenrollment through the lens of student satisfaction with the projection of Teaching, Social and Cognitive Presence. In comparing the highest and lowest disenrollment quartiles of all courses at American Public University the

Rank	Year	Title	Source	Type	Content
					value of effective Instructional Design and Organization, and initiation of the Triggering Event phase of Cognitive Presence were found to be significant predictors of student satisfaction in the lowest disenrollment quartile. For the highest disenrollment quartile, the lack of follow-through vis-a-vis Facilitation of Discourse and Cognitive Integration were found to be negative predictors of student satisfaction. (Contains 5 tables and 1 figure.) Keywords: Online Courses Electronic Learning

Rank	Year	Title	Source	Type	Content
					Dropouts School Holding Power Learner Engagement College Students Satisfaction Student Attitudes Predictor Variables
2	2017	Enhancing Cognitive Presence in Online Case Discussions with	American Journal of Distance Education.	Journal Articles.	The researchers in this study examined the influence of questions designed with the Practical Inquiry Model (PIM), compared with the regular (playground) questions, on students' levels of cognitive presence in online discussions. Students' discussion postings

Rank	Year	Title	Source	Type	Content
		Questions Based on the Practical Inquiry Model.			were collected and categorized according to the four levels of cognitive presence: "triggering events," "exploration," "integration," and "resolution." The data were analyzed using quantitative content analysis and nonparametric statistics. Results revealed that students' responses to questions based on the PIM resulted in higher levels of students' cognitive presence--integration of ideas and resolution of problems--compared with the responses based on the regular (playground) questions. These results suggest that instructors can use the PIM as a guiding

Rank	Year	Title	Source	Type	Content
					framework to design questions that may influence cognitive presence in online discussions. Keywords: Graduate Students Masters Programs Cognitive Processes Web Based Instruction Online Courses Case Method (Teaching Technique) Communities of Practice Computer Mediated Communication Questioning Techniques

Rank	Year	Title	Source	Type	Content
					Inquiry Models Content Analysis Nonparametric Statistics Problem Solving Comparative Analysis Instructional Design Prompting Interrater Reliability Coding Statistical Analysis
3	2017	Integrating the SOP [superscr	Educational Technology & Society.	Journal Articles.	This study explored student teachers' cognitive presence and learning achievements by integrating the SOP [superscript

Rank	Year	Title	Source	Type	Content
		ipt 2] Model into the Flipped Classroom to Foster Cognitive Presence and Learning Achievements.			2] Model in which self-study (S), online group discussion (O) and double-stage presentations (P[superscript 2]) were implemented in the flipped classroom. The research was conducted at a university in Taiwan with 31 student teachers. Pre- and post-worksheets measuring knowledge of educational issues were administered before and after group discussion. Quantitative content analysis and behavior sequential analysis were used to evaluate cognitive presence, while a paired-samples t-test analyzed learning achievement. The results showed that the

Rank	Year	Title	Source	Type	Content
					participants had the highest proportion of "Exploration," the second largest rate of "Integration," but rarely reached "Resolution." The participants' achievements were greatly enhanced using the SOP [superscript 2] Model in terms of the scores of the pre- and post-worksheets. Moreover, the groups with a higher proportion of "Integration" (I) and "Resolution" (R) performed best in the post-worksheets and were also the most progressive groups. Both high- and low-rated groups had significant correlations between the "I" and "R" phases, with "I" [right arrow] "R" in the

Rank	Year	Title	Source	Type	Content
					low-rated groups but "R" [right arrow] "I" in the high-rated groups. The instructional design of the SOP [superscript 2] Model can be a reference for future pedagogical implementations in the higher educational context. Keywords: Educational Technology Technology Uses in Education Group Discussion Computer Mediated Communication Homework Video Technology College Students

Rank	Year	Title	Source	Type	Content
					Worksheets Pretests Posttests Statistical Analysis Achievement Correlation Foreign Countries Scoring Rubrics Teaching Methods Student Behavior Knowledge Level Interpersonal Relationship
4	2013	Technology Readiness as a	ProQuest LLC.	Dissertations/Theses - Doctoral	Online education depends on a variety of technology tools for cognitive-related activities; however it is unclear whether the

Rank	Year	Title	Source	Type	Content
		Predictor of Cognitive Presence in Online Higher Education.		Dissertations.	current proliferation of tools is an indicator of a learner's readiness to use them effectively to meet learning objectives is unclear. Because the effectiveness of the online experience is a measure of a learner's perception of cognitive presence--the extent to which members of a technology-based community of inquiry are able to construct meaning--the purpose of this study was to explain the possible relationship between the learner's technology readiness and cognitive presence. Using the community of inquiry framework and the technology readiness index, a sample of 88 online higher education students

Rank	Year	Title	Source	Type	Content
					answered an online survey. The main research question queried if technology readiness was predictive of cognitive presence in online higher education. Use of linear regression indicated that the technology readiness sub constructs of "optimism" and "innovativeness" were significant predictors of cognitive presence but "discomfort" and "insecurity" had no predictive effect. Additionally, "insecurity" predicted the "triggering event," "optimism" predicted "exploration," and "discomfort" predicted "resolution." Incorporating the study's findings

Rank	Year	Title	Source	Type	Content
					into the development, administration, and instruction of online courses may effect positive social change by enhancing students' technology readiness and cognitive presence in terms of epistemic engagement within a technology-based community of inquiry. [The dissertation citations contained here are published with the permission of ProQuest LLC. Further reproduction is prohibited without permission. Copies of dissertations may be obtained by Telephone (800) 1-800-521-0600. Web page: http://www.proquest.com/en-US/products/dissertations/individuals.shtml.]

Rank	Year	Title	Source	Type	Content
					Keywords: Online Courses Educational Technology Learner Engagement Correlation Readiness College Students Online Surveys Predictor Variables Regression (Statistics) Student Characteristics Cognitive Processes
5	2017	Creating a	Journal of Library &	Journal Articles.	According to the Community of Inquiry (CoI) model (Garrison,

Rank	Year	Title	Source	Type	Content
		Community of Inquiry in Online Library Instruction.	Information Services In Distance Learning.		Anderson, & Archer, 2000), an enriching educational experience online in a collaborative learning environment requires three interdependent elements: social presence, teaching presence, and cognitive presence. Social presence provides interaction in the online environment that allows students to feel like they are in a supportive and open environment. Teaching presence refers not just to teacher-student interaction during the lesson or course duration, but also to a teacher's ability to design an effective learning environment. Cognitive presence in the CoI model is knowledge generated

Rank	Year	Title	Source	Type	Content
					from collaborative interaction. This model has been well-studied in the literature, and has been shown to be a meaningful framework for course development. However, more exploration of CoI in relation to library distance instruction is needed. This article describes the Community of Inquiry model and provides information about the three presences and how they can improve online educational environments. Keywords: Communities of Practice Web Based Instruction Library Instruction

Rank	Year	Title	Source	Type	Content
					Online Courses Models Distance Education Teacher Role Teacher Student Relationship Instructional Design Curriculum Development Asynchronous Communication Computer Mediated Communication Video conferencing
6	2007	Creating Shared Understa	Internet and Higher Education.	Journal Articles.	This study investigated the process by which shared understanding develops in a chat

Rank	Year	Title	Source	Type	Content
		nding through Chats in a Community of Inquiry.			learning space. It used a practical inquiry model to assess the development of cognitive presence. The study also explored how the pattern of conversation in synchronous discussion supports cognitive presence and how cognitive presence changes over time. Results show that there is a pattern among group members that involves reacquainting themselves through social presence and orienting themselves to the cognitive task through teaching presence. Individual meaning contributed by each member of the group through triggering events and

Rank	Year	Title	Source	Type	Content
					exploratory statements is transformed as members see the text on the screen and respond to it through questioning and collective exploration. This group exploration enables the transition to shared understanding. Keywords: Computer Mediated Communication Inquiry Group Discussion Computer Uses in Education Group Dynamics Educational Technology Technology Integration

Rank	Year	Title	Source	Type	Content
7	2005	Creating Cognitive Presence in a Blended Faculty Development Community.	Internet and Higher Education.	Journal Articles.	The focus of this study was to understand how a blended learning approach can support the inquiry process (cognitive presence) in a faculty development context. The findings from this study indicate that there are several key differences and similarities in cognitive presence between face-to-face and online discussions. These differences and similarities are specifically related to the four phases of cognitive presence of the practical inquiry model. A comparison of the face-to-face and online discussion forums indicates that: a slightly higher

Rank	Year	Title	Source	Type	Content
					percentage of "triggering events" occurred in the face-to-face discussions; "exploration" was the dominant phase in both environments; a noticeably greater percentage of comments were coded for "integration" in the online discussions; and the "resolution/application" phase was almost non-existent in both forms of discussion. The results from this study imply that an increased emphasis should be placed on teaching presence within a blended learning environment to ensure that participants achieve resolution in the inquiry cycle. (Contains 4 tables.)

Rank	Year	Title	Source	Type	Content
					Keywords: Computer Mediated Communication Faculty Development Online Courses Comparative Analysis Educational Environment
8	2013	Cognitive Presence in a Virtual Learning Community: An	Journal of Distance Education.	Journal Articles.	This study aimed to investigate the existence of cognitive presence as one of the elements of the Community of Inquiry framework in virtual centers for undergraduate students of science and technology. To achieve the purpose of this study, first a questionnaire was

Rank	Year	Title	Source	Type	Content
		EFL Case.			uniquely developed on the basis of the suggestions made in the literature reviewing the indicators of cognitive element. The questionnaire was then administered to undergraduate students ("N" = 107) who were studying a technical or a technological course in the Iran University of Science and Technology and Khajeh Nasir Toosi University of Technology. Analysis of the questionnaire data showed that (a) the "Exploration" and "Resolution" categories appeared more frequently than others in the virtual centers of this study and (b) the indicators of

Rank	Year	Title	Source	Type	Content
					"Divergence," "Information Exchange," and "Applying New Ideas" were hierarchically frequent. In order to promote and sustain cognitive presence, this study recommends that virtual language educators incorporate the indicators of cognitive presence into the online learning environment. Keywords: Undergraduate Students Communities of Practice Science Education Technology Education Questionnaires Foreign Countries

Rank	Year	Title	Source	Type	Content
					Electronic Learning Thinking Skills Cognitive Processes Statistical Significance Virtual Classrooms Statistical Analysis Correlation English (Second Language) Technology Uses in Education Student Participation Learner Engagement
9	2012	The Effects of Collabor	ProQuest LLC.	Dissertations/Theses - Doctoral	This study continues the exploration of the Community of Inquiry framework and how collaborative technologies,

Rank	Year	Title	Source	Type	Content
		ative Tools on Student Perceptions of the Community of Inquiry Framework in an Online Course.		Dissertations.	specifically wikis, can be used to impact student perception of social presence. The subjects were 78 graduate education students in three differently contrived sections of the same online course. Participants completed the Community of Inquiry Survey at the end of the term, which measured their perceived level of teaching, social, and cognitive presence during the course. The experimental setting utilized a single instructor teaching one course, and randomly assigned students. Each section had students collaborate using a different tool (synchronous wiki,

Rank	Year	Title	Source	Type	Content
					asynchronous wiki, and discussion board-only). All subjects perceived high levels of the three presences when compared with previous studies. Students collaborating using an asynchronous wiki perceived significantly more social presence than those using only a discussion board. Specifically, students perceived greater levels of trust and group cohesion when the course design incorporated a wiki for small group collaboration. [The dissertation citations contained here are published with the permission of ProQuest LLC. Further reproduction is prohibited

Rank	Year	Title	Source	Type	Content
					without permission. Copies of dissertations may be obtained by Telephone (800) 1-800-521-0600. Web page: http://www.proquest.com/en-US/products/dissertations/individuals.shtml.] Keywords: Online Courses Student Attitudes Graduate Students Electronic Publishing Cooperation Computer Mediated Communication Trust (Psychology)

Rank	Year	Title	Source	Type	Content
					Group Dynamics Course Descriptions Teaching Methods Inquiry
10	2017	Three Interaction Patterns on Asynchronous Online Discussion Behaviours: A Methodo	Journal of Computer Assisted Learning.	Journal Articles.	An asynchronous online discussion (AOD) is one format of instructional methods that facilitate student-centered learning. In the wealth of AOD research, this study evaluated how students' behavior on AOD influences their academic outcomes. This case study compared the differential analytic methods including web log mining, social network analysis and content analysis

Rank	Year	Title	Source	Type	Content
		logical Comparison.			which were selected by three interaction patterns: person to system (P2S), person to person (P2P) and person to content (P2C) interaction. Forty-three undergraduate students participated in an online discussion forum for 12 weeks. Multiple regression analyses with the predictor variables from P2S, P2P and P2C and with a criterion variable of a final grade indicated several interesting findings. For P2S analysis, visits on board (VOB) had a significant variable to predict final grades. Also, the result of P2P analysis proved that in-degree and out-degree centrality predicted final

Rank	Year	Title	Source	Type	Content
					grades. The P2C results based on cognitive presence represent that students' messages were mostly affiliated to the exploration and integration levels and also predicted the final grades. This study ultimately demonstrated the effectiveness of using multiple analytic methodologies to address and facilitate students' participation at AOD. Keywords: Computer Mediated Communication Interaction Process Analysis Undergraduate Students Predictor Variables

Rank	Year	Title	Source	Type	Content
					Multiple Regression Analysis Social Networks Content Analysis Grades (Scholastic) Teaching Methods Student Centered Learning Case Studies Comparative Analysis Behavior Patterns Student Behavior Network Analysis Group Discussion

Retrieval Ranking List 3:

Rank	Year	Title	Source	Type	Content
1	2018	Cognitive Presence in Peer Facilitated Asynchronous Online Discussion: The Patterns and How to Facilitate.	ProQuest LLC.	Dissertations/Theses - Doctoral Dissertations.	This study, in the context of peer-facilitated asynchronous online discussion, explored the characteristics and patterns of students' cognitive presence, and examined the practices that aim to enhance cognitive presence development. Participants were 53 students from a graduate-level online course that focused on the integration of educational technologies. Data were collected from discussion transcripts, student survey, student artifacts, and researcher's observations. Results demonstrated four phases of students' cognitive presence: Triggering event, Exploration,

Rank	Yea r	Title	Source	Type	Content
					Integration, and Resolution. Among the four phases, students' cognitive presence tended to aggregate at the middle phases: Integration and Exploration. Percentage of the Resolution was very low. The distribution of students' discussion behaviors further revealed: a) the hierarchical relationship between the four phases: Integration and Resolution involved a higher-level of cognitive engagement, and Triggering event and Exploration involved a lower-level of cognitive engagement; b) the phase of Resolution heavily relied on experiment, while the

Rank	Yea r	Title	Source	Type	Content
					other three phases heavily relied on making use of personal experience; c) creating of cognitive presence occurred in both the private space of individual activities and the shared space of having dialogues. The conversation analysis of threads and episodes explored the temporal evolvement of cognitive presence. The results showed that, in an ongoing discussion, students' cognitive presence evolved in a non-linear way, rather than strictly phase by phase as suggested by the PI model. Experiments were designed and conducted to

Rank	Yea r	Title	Source	Type	Content
					determine the effects of two pedagogical interventions -- 1) providing guidance on peer facilitation techniques; 2) asking students to label their posts. The results showed that the Intervention 1 and the combination of two interventions credibly improved students' cognitive presence. They were especially effective in improving Integration, a higher level of cognitive presence. After having added Intervention 2, cognitive presence increased from the first-half to the second-half semester, although the improvement was not found to be statistically

Rank	Yea r	Title	Source	Type	Content
					credible. This study confirmed the close association between and among cognitive presence, social interaction, and peer facilitation. The results clearly showed that Intervention 1 -- providing guidance on peer facilitation credibly improved students' social interaction and peer facilitation. However, Mixed findings were obtained for Intervention 2 -- asking students to label their posts. It was found that Intervention 2 positively increased students' social interaction. However, it did not show any impact on students' peer facilitation behaviors. It is

Rank	Yea r	Title	Source	Type	Content
					also worth noting that the effect of the combination of two interventions was much larger than any single one of them. Conversation analysis was conducted to zoom in on the dynamic process of discussion. The cases revealed that when students were provided with the guidance on peer facilitation techniques, they tended to use a variety of facilitation techniques in a strategic way to help peers to achieve a sustained and deeper-level conversation. Compared to the control group, the students in the treatment group showed more peer facilitation behaviors, which

Rank	Yea r	Title	Source	Type	Content
					led to more conversations and more higher-level cognitive presence. This study has unpacked the complexity of students' cognitive presence in a peer-facilitated discussion environment, especially when students are coached in performing teaching presence. The results shed light on the pedagogical practices and strategies of creating an online learning community that incubates rich cognitive presence. Finally, implications are discussed for the research and practices in online instruction and discussion analytics. [The

Rank	Year	Title	Source	Type	Content
					dissertation citations contained here are published with the permission of ProQuest LLC. Further reproduction is prohibited without permission. Copies of dissertations may be obtained by Telephone (800) 1-800-521-0600. Web page: http://www.proquest.com/en-US/products/dissertations/individuals.shtml.] Keywords: Online Courses Graduate Students Asynchronous Communication Computer Mediated Communication

Rank	Year	Title	Source	Type	Content
					Discussion (Teaching Technique) Correlation Cognitive Processes Learner Engagement Intervention Program Effectiveness Peer Teaching Peer Influence Facilitators (Individuals)
2	2011	An Exploration of Differences	Journal of Asynchronous Learning	Journal Articles.	Though online enrollments continue to accelerate at a rapid pace, there is significant concern over student retention. With drop

Rank	Year	Title	Source	Type	Content
		between Community of Inquiry Indicators in Low and High Disenrollment Online Courses.	Networks.		rates significantly higher than in face-to-face classes it is imperative that online providers develop an understanding of factors that lead students to disenroll. This study utilizes a data mining approach to examine course-level disenrollment through the lens of student satisfaction with the projection of Teaching, Social and Cognitive Presence. In comparing the highest and lowest disenrollment quartiles of all courses at American Public University the value of effective Instructional Design and Organization, and initiation of the Triggering Event

Rank	Yea r	Title	Source	Type	Content
					phase of Cognitive Presence were found to be significant predictors of student satisfaction in the lowest disenrollment quartile. For the highest disenrollment quartile, the lack of follow-through vis-a-vis Facilitation of Discourse and Cognitive Integration were found to be negative predictors of student satisfaction. (Contains 5 tables and 1 figure.) Keywords: Online Courses Electronic Learning Dropouts School Holding Power

Rank	Yea r	Title	Source	Type	Content
					Learner Engagement College Students Satisfaction Student Attitudes Predictor Variables
3	2014	Reflection as an Indicator of Cognitive Presence.	E-Learning and Digital Media.	Journal Articles.	In the Community of Inquiry (CoI) model, cognitive presence indicators can be used to evaluate the quality of inquiry in a discussion forum. Engagement in critical thinking and deep knowledge can occur through reflective processes. When learners move through the four phases of cognitive presence (triggering, exploration,

Rank	Yea r	Title	Source	Type	Content
					integration, resolution), the processes of discussion and reflection are important in developing deep understanding. In this article, data from the online discussion archives within a blended teacher-education course are analysed using the cognitive presence indicators from the CoI with the additional indicator of reflection. This study indicates that when instructors structure online discussions appropriately, learners are able to share and document their thinking and reflect on their contributions and the perspectives of others while

Rank	Year	Title	Source	Type	Content
					developing new or deeper knowledge. To facilitate the coding of reflective activities and online posts the researcher proposes modifying the resolution phase of the original cognitive presence coding protocol to include an additional reflection indicator. Keywords: Communities of Practice Electronic Learning Discussion Groups Archives Blended Learning Teacher Education

Rank	Year	Title	Source	Type	Content
					Learning Processes Cognitive Processes Reflection Coding Shared Resources and Services
4	2013	Technology Readiness as a Predictor of Cognitive Presence in Online Higher Education.	ProQuest LLC.	Dissertations/Theses - Doctoral Dissertations.	Online education depends on a variety of technology tools for cognitive-related activities; however it is unclear whether the current proliferation of tools is an indicator of a learner's readiness to use them effectively to meet learning objectives is unclear. Because the effectiveness of the online experience is a measure of a learner's perception of cognitive

Rank	Yea r	Title	Source	Type	Content
					presence--the extent to which members of a technology-based community of inquiry are able to construct meaning--the purpose of this study was to explain the possible relationship between the learner's technology readiness and cognitive presence. Using the community of inquiry framework and the technology readiness index, a sample of 88 online higher education students answered an online survey. The main research question queried if technology readiness was predictive of cognitive presence in online higher education. Use of linear regression indicated that

Rank	Yea r	Title	Source	Type	Content
					the technology readiness sub constructs of "optimism" and "innovativeness" were significant predictors of cognitive presence but "discomfort" and "insecurity" had no predictive effect. Additionally, "insecurity" predicted the "triggering event," "optimism" predicted "exploration," and "discomfort" predicted "resolution." Incorporating the study's findings into the development, administration, and instruction of online courses may effect positive social change by enhancing students' technology readiness and cognitive presence

Rank	Yea r	Title	Source	Type	Content
					in terms of epistemic engagement within a technology-based community of inquiry. [The dissertation citations contained here are published with the permission of ProQuest LLC. Further reproduction is prohibited without permission. Copies of dissertations may be obtained by Telephone (800) 1-800-521-0600. Web page: http://www.proquest.com/en-US/products/dissertations/individuals.shtml.] Keywords: Online Courses Educational Technology

Rank	Year	Title	Source	Type	Content
					Learner Engagement Correlation Readiness College Students Online Surveys Predictor Variables Regression (Statistics) Student Characteristics Cognitive Processes
5	2011	Cognitive Presence in Asynchronous Online Learning: A	Journal of Computer Assisted Learning.	Journal Articles.	Some scholars argue that students do not achieve higher level learning, or cognitive presence, in online courses. Online discussion has been proposed to

Rank	Year	Title	Source	Type	Content
		Comparison of Four Discussion Strategies.			bridge this gap between online and face-to-face learning environments. However, the literature indicates that the conventional approach to online discussion--asking probing questions--does not necessarily advance the discussion through the phases of cognitive presence: triggering events, exploration, integration and resolution, which are crucial for deep knowledge construction. Using mixed methods, we examined the contribution of four scenario-based online discussion strategies--structured, scaffolded, debate and role play--to the

Rank	Yea r	Title	Source	Type	Content
					learners' cognitive presence, the outcome of the discussion. Learners' discussion postings within each strategy were segmented and categorized according to the four phases. The discussion strategies, each using the same authentic scenario, were then compared in terms of the number of segments representing these phases. We found that the structured strategy, while highly associated with triggering events, produced no discussion pertaining to the resolution phase. The scaffolded strategy, on the other hand, showed a strong association with the

Rank	Yea r	Title	Source	Type	Content
					resolution phase. The debate and role-play strategies were highly associated with exploration and integration phases. We concluded that discussion strategies requiring learners to take a perspective in an authentic scenario facilitate cognitive presence, and thus critical thinking and higher levels of learning. We suggest a heuristic for sequencing a series of discussion forums and recommend areas for further related research. Keywords: Electronic Learning

Rank	Year	Title	Source	Type	Content
					Computer Mediated Communication Online Courses Asynchronous Communication Discussion (Teaching Technique) Distance Education Comparative Analysis Learning Strategies Vignettes Debate Scaffolding (Teaching Technique) Role Playing

Rank	Year	Title	Source	Type	Content
					Critical Thinking Thinking Skills Heuristics
6	2017	Using the Community of Inquiry Framework to Scaffold Online Tutoring.	International Review of Research in Open and Distributed Learning.	Journal Articles.	Tutoring involves providing learners with a suitable level of structure and guidance to support their learning. This study reports on an exploration of how to design such structure and guidance (i.e., learning scaffolds) in the Chinese online educational context, and in so doing, answer the following two questions: (a) What scaffolding strategies are needed to design online tutoring, and (b) How should different

Rank	Yea r	Title	Source	Type	Content
					levels of scaffolding intensity be emphasized in different stages of online tutoring in such educational contexts? A model for online tutoring using the Community of Inquiry framework was developed and implemented in this study. It focused attention on both the critical role of the tutor in online learning and the importance of scaffolding in online tutoring. Both qualitative and quantitative methods were used to collect data, including questionnaires, interviews, and content analysis. In considering the variation of scaffolding throughout the online

Rank	Yea r	Title	Source	Type	Content
					course, results showed that: (a) As long as a high degree of social presence is established in the initial phase, scaffolds for social presence can be withdrawn gradually throughout the course; (b) High-intensity teaching presence is much more important in the mid-phase of the course than in other phases; (c) "Discourse facilitation" should be emphasized for teaching presence in the mid-phase, while "direct instruction" scaffolding is needed in the last phase; and (d) The greatest need for scaffolding of cognitive presence occurs in the final phase of the course.

Rank	Year	Title	Source	Type	Content
					Keywords: Online tutoring Online presence Scaffolding Community of inquiry
7	2014	Learning to Argue in a Connected World: The Arc of Productive Disciplinary Engagement in a High School	ProQuest LLC.	Dissertations/Theses - Doctoral Dissertations.	Calls to virtually break down school walls through connected and blended learning environments are ubiquitous as of late as technologies in service of learning evolve and as schools are under pressure to change. Within the subject area of English Language Arts, there is a dearth of research or information on how to facilitate these new,

Rank	Yea r	Title	Source	Type	Content
		Academic Social Network.			digitally enhanced methods in high schools in a way that approaches or leads to productive disciplinary engagement (PDE). The current study describes one such scenario in which an academic social network (Remix) was used for the retrieval of curriculum, the storage of student work, and the exchange for both social and academic conversations with an entire freshman class of high school students across three teachers and eight classrooms. The seven-week curriculum focused on learning to read, analyze, and write evidence-based, classical

Rank	Yea r	Title	Source	Type	Content
					arguments. Experts in argumentation (e.g., lawyers, journalists, grant writers, ex-English teachers, etc.) interacted with the students three times during the arc of the unit to give targeted feedback to students during their growing understanding of argumentation. To determine the degree to which PDE occurred within the platform, the posts of twenty-five randomly selected students, who had at least one interaction with an expert, were downloaded and coded for Social and Cognitive Presence--two domains of the Community of Inquiry Model.

Rank	Yea r	Title	Source	Type	Content
					The analysis illustrated that Social Presence acted as connective tissue to academic tasks and that socializing moved to an academic orientation as students collaborated and worked toward a common goal. Cognitive Presence also moved from trigger events which included recognizing and puzzling over contemporary issues to the exploration and integration of ideas as the unit progressed. The discourses associated with academic social networks proved slightly troublesome for students, lending credence that they need more

Rank	Yea r	Title	Source	Type	Content
					practice in such platforms when posting and responding to academic content. A second investigation was completed to look more specifically at expert feedback in relation to PDE components, argumentation, and curricular activity. Differences between the three feedback interactions proved scientifically significant, thus illustrating experts adjusted their responses to students depending on the task. Experts moved from problematizing student arguments at the trigger stage of topic selection to directing students as to how to fix their

Rank	Year	Title	Source	Type	Content
					arguments during the integration stage of production, thus holding students accountable to disciplinary norms as the unit progressed. Advice for how to utilize a social network and work with outside experts is also covered. [The dissertation citations contained here are published with the permission of ProQuest LLC. Further reproduction is prohibited without permission. Copies of dissertations may be obtained by Telephone (800) 1-800-521-0600. Web page: http://www.proquest.com/en-US/products/dissertations/individuals.shtml.]

Rank	Yea r	Title	Source	Type	Content
					Keywords: Social Networks High School Freshmen English Language Arts Blended Learning Secondary School Teachers Curriculum Feedback (Response) Persuasive Discourse Specialists Cooperative Learning Information Technology

Rank	Year	Title	Source	Type	Content
8	2015	Literacy Co-Teaching with Multi-Level Texts in an Inclusive Middle Grade Humanities Class: A Teacher-Researcher Collaboration.	Journal of Inquiry and Action in Education.	Journal Articles.	This article reports on a middle school literacy intervention implemented during a yearlong teacher-researcher collaboration. The purpose of this collaboration was to combine and adjust commonly recommended pedagogical approaches to address the literacy needs of a heterogeneous group of seventh graders attending an urban school. University researchers designed and implemented the intervention with an interdisciplinary team of three teachers. The intervention drew on sociocultural theories of language and learning. It had

Rank	Yea r	Title	Source	Type	Content
					three main features: integration of English and social studies, multi-level texts, and co-teaching of heterogeneous groups. Qualitative data included field notes from classroom observations and planning meetings, transcripts from teacher interviews, and classroom artifacts. Data were analyzed as they were collected and used in planning sessions. Additional analysis after the intervention ended focused on exploration of critical events reflecting convergence and divergence of teachers' and researchers' perspectives on the

Rank	Yea r	Title	Source	Type	Content
					intervention features. Findings were organized around three representative critical events, one per intervention feature. Implications of results for future middle grade co-teaching literacy interventions were explored. Keywords: Team Teaching Literacy Education Middle School Teachers Educational Research Partnerships in Education Humanities Instruction Intervention

Rank	Year	Title	Source	Type	Content
					Urban Schools Grade 7 Social Studies Language Arts Special Education Interdisciplinary Approach Reading Materials Heterogeneous Grouping Qualitative Research Affordances Observation Semi Structured Interviews

Rank	Year	Title	Source	Type	Content
9	2007	Creating Shared Understanding through Chats in a Community of Inquiry.	Internet and Higher Education.	Journal Articles.	This study investigated the process by which shared understanding develops in a chat learning space. It used a practical inquiry model to assess the development of cognitive presence. The study also explored how the pattern of conversation in synchronous discussion supports cognitive presence and how cognitive presence changes over time. Results show that there is a pattern among group members that involves reacquainting themselves through social presence and orienting themselves to the cognitive task through teaching

Rank	Yea r	Title	Source	Type	Content
					presence. Individual meaning contributed by each member of the group through triggering events and exploratory statements is transformed as members see the text on the screen and respond to it through questioning and collective exploration. This group exploration enables the transition to shared understanding. Keywords: Computer Mediated Communication Inquiry Group Discussion Computer Uses in Education

Rank	Year	Title	Source	Type	Content
					Group Dynamics Educational Technology Technology Integration
10	2009	Exploration of an E-Learning Model to Foster Critical Thinking on Basic Science Concepts during Work Placements.	Computers & Education.	Journal Articles.	We designed an e-learning model to promote critical thinking about basic science topics in online communities of students during work placements in higher education. To determine the effectiveness and efficiency of the model we explored the online discussions in two case studies. We evaluated the quantity of the interactions by looking at quantitative data of the discussion "threads" and we

Rank	Yea r	Title	Source	Type	Content
					evaluated the quality of the discussion by content analysis of the individual messages. Both the procedural facilitation of the discussion and the instrument for content analysis were based on Garrison's "Practical Inquiry model of Cognitive Presence". Furthermore, we explored the experiences of the students and moderators by interviewing them and we organized their perceptions using the framework of an activity system. On the basis of the quantitative and qualitative data we conclude that the e-learning model was successful in establishing a

Rank	Yea r	Title	Source	Type	Content
					dialogue among a group of students and an expert during work placements at different locations. The "Practical Inquiry model" was useful in facilitating a sustained on-topic discourse involving critical thinking. Although the amount of critical thinking was moderate, the results suggest ways to increase integration and resolution activities in the online discussions. (Contains 2 tables and 7 figures.) Keywords: Critical Thinking Content Analysis

Rank	Year	Title	Source	Type	Content
					Electronic Learning Models Scientific Concepts Job Placement Higher Education Case Studies Discussion Groups Evaluation Interviews Technology Uses in Education Educational Technology