BILINGUAL PUBLISHING CO.
Pioneer of Global Academics Since 1984

Journal of Computer Science Research
http://ojs.bilpublishing.com/index.php/jcsr

ARTICLE

An Ontology-based Ranking Model in Search Engines

Yu Hou* **Lixin Tao**

Pace University, New York, United States

ARTICLE INFO

Article history
Received: 28 June 2019
Accepted: 23 July 2019
Published Online: 30 July 2019

Keywords:
Search engine
Knowledge representation
Ontology
Semantic similarity
Page ranking

ABSTRACT

As the tsunami of data has emerged, search engines have become the most powerful tool for obtaining scattered information on the internet. The traditional search engines return the organized results by using ranking algorithm such as term frequency, link analysis (PageRank algorithm and HITS algorithm) etc. However, these algorithms must combine the keyword frequency to determine the relevance between user's query and the data in the computer system or internet. Moreover, we expect the search engines could understand users' searching by content meanings rather than literal strings. Semantic Web is an intelligent network and it could understand human's language more semantically and make the communication easier between human and computers. But, the current technology for the semantic search is hard to apply. Because some meta data should be annotated to each web pages, then the search engine will have the ability to understand the users intend. However, annotate every web page is very time-consuming and leads to inefficiency. So, this study designed an ontology-based approach to improve the current traditional keyword-based search and emulate the effects of semantic search. And let the search engine can understand users more semantically when it gets the knowledge.

1. Introduction

With the massive data on Internet, the increasing information has been available to the users. However, it is hard for the users to search satisfying information when they are facing the large amount information. The search engine has become the most powerful tool for obtaining scattered information on the Internet. A search engine is a system that is designed to search for information stored on a computer system or the World Wide Web. The search results are generated by the ranking algorithm, and then present to the users in a specific order. In general, the search engine uses this algorithm to determine the relevance and weight of the results to search queries (or keywords). The traditional ranking algorithm is based on the term frequency. Which means the algorithm determines the relevance between the query and the document is based on the frequency that the query appears in the document. However, there is a limitation when a web page has a high term frequency but low quality, such as advertisement web page. Then, Link Analysis is used to filter out these web pages. The most recent and most refined algorithms are PageRank [1] and HITS [2]. As all the methods I listed above, they determine the relevance between the query and document is based on the term (Link Analysis just filter out the low-quality web pages). To improve users' search experience and assist them to achieve more useful and accurate result has become an important

**Corresponding Author:*
Yu Hou,
Pace University, New York, United States;
Email: yh50276p@pace.edu

DOI: https://doi.org/10.30564/jcsr.v1i2.972

challenge for World Wide Web Consortium (W3C). The semantic web was proposed by Berners-Lee [3], and he described the semantic web as a component of "Web 3.0". Basically, the semantic web is an intelligent network that not only understands human language, but also make the communication between human and computers easier. It recognizes and processes the users' retrieval request from the semantic level. By annotating the resource objects in the network and processing the users' query expression semantically, the natural language contains a semantic logical relationship. Then the semantic search engine can execute the extensive and effective semantic reasoning in the network environment, so as to realize users' retrieval more accurately and comprehensively. This is the most common solution for the Semantic Web at present [4].

Current technology for the semantic search is hard to apply. Because it is inefficiency for the search engine if it annotate every web pages. Therefore, in this paper, we proposed an Ontology-based Web Pages Ranking Model to improve the current traditional keyword-based search and emulate the effects of semantic search. The model will rank the web pages based on the relevance between the keyword and the web pages by introducing ontology. As we expected, the proposed model could not only consider the semantic similarity in the ontology, but also consider the structural factors of the concept in the ontology, so as to improve the users' search experience on a semantic level. Therefore, the ontology is introduced to describe domain knowledge.

The main contributions of this paper are, 1. convert a symbolic and logical system (knowledge, in this paper is ontology) into a machine-readable quantification result. Because it is difficult for the applications to involve numerical computing in continuous spaces. 2. reuse the acquired domain knowledge efficiently through the methods provided in this study. 3. calculate the degree of association of concepts in the domain knowledge. In the rest of the paper, the related work of the semantic search will be introduced in section 2, and the model details will be illustrated in section 3. Then we will evaluate the model in section 4. Finally, we will make a conclusion of this research.

2. Related Work

Ranking algorithm plays an important role to determine the relevance and weight of the results to search queries (or keywords) in a search engine. The traditional search engines- the widest utilized search methods, has several ranking methods such as Keyword-based web pages ranking and link analysis (PageRank algorithm and HITS algorithm). The PageRank algorithm [5] assigns the same score to each page initially, then updating the PageRank score for each page by the iterative recursive calculation until the score is stable. The HITS algorithm [2] is also a basic and important algorithm in link analyses, and it has been implemented by the Teoma search engine. There are two basic definitions in the HITS algorithm which are "Authority page" and "Hub page". The "Authority" page refers to high-quality web pages related to a certain 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' query in the massive web pages.

Obviously, these algorithms must combine the keyword-based algorithm to optimize the ranking of web pages. As we mentioned previously, we expect the search engines could understand users' searching by content meanings rather than literal strings. Therefore, the semantic search has been introduced. It could be a highly efficient search engine that retrieves the web pages by considering the most relevant information. Ontology [6] is one of the most important concepts used in the semantic web infrastructure, and RDF(S) (Resource Description Framework/Schema) and OWL (Web Ontology Languages) are two W3C recommended data representation models to represent ontologies. Since Tim Berners-Lee proposed the Semantic Web [3], many researchers have attempted to apply domain ontology to information retrieval. Liyang Yu [4] marked the web pages by using the created domain knowledge base, namely, adding some extra data or information on the web page to describe some specific characteristics of the web pages, and the data come from the domain knowledge base. Then, the search engine collects the information via the crawler and create an index table. When the crawler reaches a web page which has not been marked up for any special semantics, the web page will add to the index table just under every index key normally. If the crawler reaches a web page which has been marked, then, it will parse and download the domain knowledge base, and add the related web pages under the index key. When a user tries to search a keyword by using the Semantic Web, the search engine will retrieve the index table and return the candidate result set under the keyword. This is a traditional semantic search engine strategy. However, annotating every web pages leads to inefficiency for the search engine. More recently, Abdelkrim Bouramoul [7] and Prerna Parmeshwaran [8] use the similarity measure to calculate the similarity between the query provided and the documents available. The distance measured in the vectorial space means the relevance be-

DOI: https://doi.org/10.30564/jcsr.v1i2.972

tween the query and the document. Then the designed the page rank model use the distance to complete the semantic rank. Anila Sahar Butt [9] proposed a DWRank model to improve the ontology search. The DWRank model is based on the Hub Score and Authority Score. The author claims that the hub Score is a measure of the centrality of a concept within an ontology. A concept is more central to an ontology if there are more relationships starting from the concept. If there is a relationship starting from the concept to another central concept, the concept is more central to an ontology. The authority score is the measure of the authoritativeness of a concept within an ontology. If there are more inter-ontology relationships ending at the ontology, an ontology is more authoritative. If there is an inter-ontology relationship starting from an authoritative ontology to the ontology, an ontology is more authoritative.

In summary, some studies use the semantic annotation to improve the performance of retrieval systems. This approach needs to mark down each web page by using domain knowledge base and add them to the index table, which is inefficiency. However, the current web data don't have the needed semantic annotations. Therefore, afterward, most studies use the semantic relationships and inference mechanism in the ontology to improve the information retrieval. However, these methods still have some limitations. For example, the semantic relationship only considers the semantic distance between concepts but ignores the structural factors of the concept in the ontology. Semantic reasoning has a higher requirement for ontology, and perfect ontology is the basis for implementing semantic reasoning. Building such domain ontology is a huge project and it is difficult to achieve.

In this paper, we proposed an ontology-based web pages ranking model by using the semantic similarity between the concepts. The semantic similarity is calculated by using the semantic relationships and structural factors between concepts in the ontology. This approach could improve the retrieval at the semantic level and alleviate the limitations listed above.

3. Methodology

Ontology is a very effective way of expressing knowledge. An ontology can be considered as a tree-structured data, so we 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, we can efficiently and accurately capture the semantic information behind the user's query and complete the semantic search. The main purpose of this study is try to how to convert a domain knowledge (ontology) into a machine-readable quantification result. Then, calculate the degree of association of concepts in the domain knowledge. Thereby improving the current traditional keyword-based search and emulate the effects of semantic search.

Based on the traditional search engine, this approach introduces the ontology to improve the search experience on the semantic level. Namely, a number of 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: 1. Create the domain knowledge base; 2. Select the candidate web pages from the index table; 3. Calculate the semantic similarity between the concepts in the ontology; 4. Score the candidate web pages by the semantic similarity and the term's TF-IDF weight; 5. Rank the candidate web pages by the score which are generated in step 4 and return the result. Figure 1 shows the framework of the model.

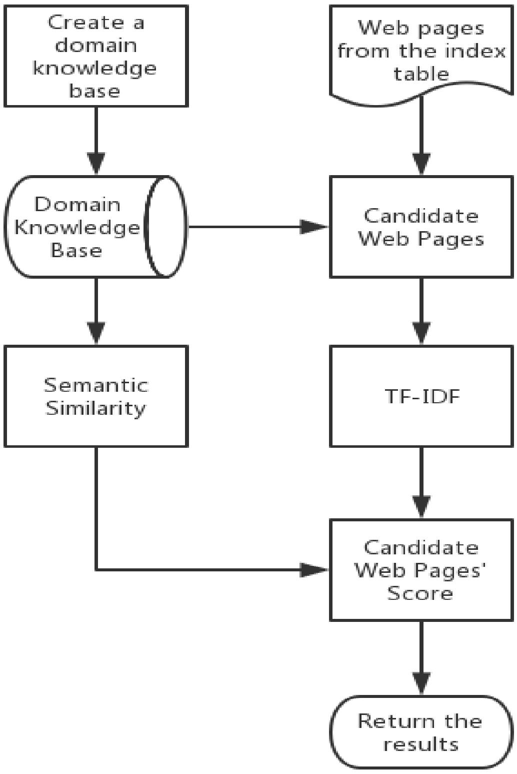

Create a domain knowledge base
Web pages from the index table
Domain Knowledge Base
Candidate Web Pages
Semantic Similarity
TF-IDF
Candidate Web Pages' Score
Return the results

Figure 1. The Framework of the Ontology-based Web Pages Ranking Model

3.1 Create the Domain Knowledge Base

Ontology originates from a philosophical concept used to describe the nature of things. Gruber from the Knowl-

DOI: https://doi.org/10.30564/jcsr.v1i2.972

edge Systems Laboratory in 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." [10] One of the reasons why ontology is important is that its consensus on the concept of a certain 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. [11] In our paper, we use Protégé (https://protege.stanford.edu/) to define the ontology and encode it, then we saved the ontology as an OWL (Web Ontology Language) file. The figure 2 is an example of ontology in a domain knowledge.

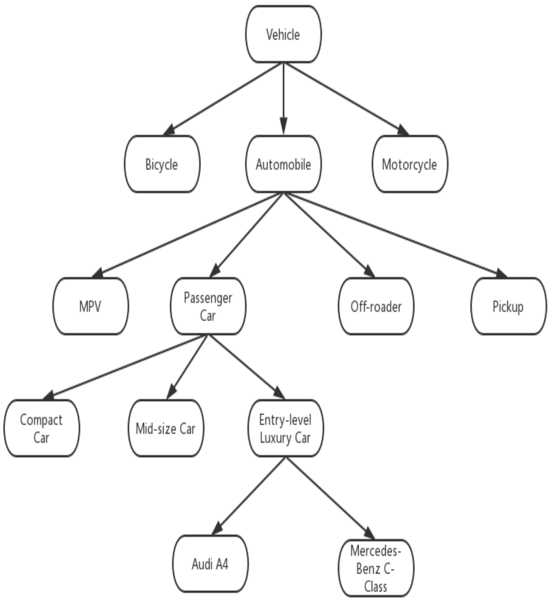


Figure 2. An example of ontology in a domain knowledge

3.2 Select Candidate Web Pages

The index table in the traditional search engine is built by every single word on the web page. [4] When we have a domain knowledge base, we will have a dictionary which is composed by the node in the ontology. The traditional keyword-based ranking method only considers the web pages from the keyword in the index table. As we know, some web pages are quite relevant with the keyword, though they do not contain the keyword. Thus, the candidate web pages in our approach are selected from every node in the ontology in the index table. We will rank the candidate web pages as the result by using the model.

3.3 Calculate the Semantic Similarity

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

(1) Semantic Distance: We can assume that X and Y are two nodes (or concept) in the ontology and the shortest path between X and Y is Semantic Distance, referred to as Dis (X, Y). The Semantic Distance is an important element when we compute the Semantic Similarity. When the distance between two conceptual paths are far, the Semantic Distance is far, and the Semantic Similarity is smaller. For example, we can calculate that the semantic distance between Audi A4 and Benz c class is 2, and the semantic distance between Audi A4 and pickup is 4; that is, the semantic similarity between Audi A4 and Benz c class is large (both are luxury car), and the Audi A4 and Pickup has a low semantic similarity (different types).

(2) 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.

$$Semantic\ Coincidence = \frac{|N(X) \cap N(Y)|}{|N(X) \cup N(Y)|} \quad (1)$$

The Semantic Coincidence represents the same degree between two concepts. For example, The semantic coincidence of Audi A4 and Benz c class is 4 over 6 (0.67), while the semantic coincidence of Audi A4 and pickup is 2 over 6 (0.34). Obviously, the semantic similarities between Audi A4 and Benz c class are higher.

(3) 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 same if they are in the different level at the ontology tree. The bigger the Level Difference is, the smaller 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 only automobile. So, the semantic similarity of the former should be greater than the latter.

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

$$Sim(X,Y) = \frac{\alpha \cdot \beta \cdot N(X) \cap N(Y)}{[Dis(X,Y)+\alpha] \cdot [|L(X)-L(Y)|+\beta] \cdot N(X) \cup N(Y)} \quad (2)$$

α and β are parameters, which can adjust the influence of three factors above. We can understand that the Sim (X, Y) has a range of (0, 1], which means all the concepts are

DOI: https://doi.org/10.30564/jcsr.v1i2.972

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.

3.4 TF-IDF

TF-IDF, short for 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. [12] 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 a very good 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 for 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. [13]

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

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (3)$$

In Equation 3, $n_{i,j}$ is the number of occurrences of the word t_i in the file d_i , and the denominator is the sum of the occurrences of all words in the file d_i. 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.

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

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:

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

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) and is rarely found in other articles (high IDF value), the word or phrase is considered to have good ability of the class distinguishing and is suitable for classification.

3.5 Score the Candidate Web Pages

As we already have a dictionary which is composed by the node in the ontology. Then we will check each candidate web page and find out what words are included in the dictionary and also appears on the web page. The model will score each candidate web pages by the semantic similarity and the term's TF-IDF weight.

As a result, we can represent our model as follows:

$$Score(keyword) = TFIDF(keyword) + \sum_i \left[TFIDE(wi) \times Sim(keyword, wi)\right] \qquad (6)$$

The w_i means the words in the dictionary except for the keyword.

3.6 Rank the Candidate Web Pages

The model will return the candidate web pages in the descending order where the score comes from the previous step. And the result will filter off the web pages when the score is 0.

The proposed model combined the TF-IDF weights with the domain knowledge base and introduces the semantic relevance of the keywords and another vocabulary in the web pages. By using the model, we can improve the users' search experience by letting the search engine achieve the capabilities to capture the conceptualizations involved in users' intention and web pages' content meaning.

4. Experiment Design

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

4.1 Data Preparation

The purpose of this study is to find a way for search engines to understand users' queries more semantically and effectively in a specific domain. This can be achieved very well by introducing ontology or/and knowledge graph. Once the search engine obtains a domain knowledge, it should reuse the knowledge flexibly through the meth-

DOI: https://doi.org/10.30564/jcsr.v1i2.972

ods we provided in this study. Therefore, the experiment should be focused on a specific field, and the data should be diversity in order to compare the performance of different models effectively.

PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), offered by the United States National Library of Medicine (NLM) which is the world's largest medical library.[14] It is a free search engine that interfaces with the Medical Literature Analysis and Retrieval System Online (MEDLINE) database which includes bibliographic information for articles from biomedical journals. And the NLM at the National Institutes of Health maintains the database. MeSH (https://www.nlm.nih.gov/mesh/meshhome.html), the medical subject headings, which is the national library of medicine's vocabulary thesaurus. It is used for indexing articles for the MEDLINE database. MeSH terms can be used to describe the article in the MEDLINE database.

The titles and abstracts of 16,105 articles were downloaded from PubMed, and then classified into three categories: Parkinson's disease, Alzheimer's disease, and Lewy Body disease according to MeSH classification.

As I described in the section 1, the traditional search engines did a lot of optimization work on the ranking algorithm (such as PageRank, HITS algorithm etc.). But these algorithms can only help search engines filter out low-quality web-pages and find out high-quality web-pages. The search engines still need to determine the degree of association between the query and the web-pages on the frequency of the query keyword appears in the web-pages. Therefore, the designed experimental prototype aims to compare that a search engine will be more semantic and effective by introduce a domain knowledge.

By designing and applying two ranking algorithms (keyword-based and ontology-based), the purpose of the experiment is to retrieve the same keyword from the prepared data (PubMed) to achieve the following assumptions:

(1) The experimental prototype can achieve higher recall rates when the ontology-based algorithm applied. Some documents or web-pages that are highly relevant to the query may not be retrieved or improved by the keyword-based algorithm. 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 basically based on literal matching as the basis for sorting. The result of no literal matching (no keyword "iPhone" in the web page) will not be searched, even if the query and the web page are semantically relevant.

(2) When the ontology-based ranking algorithm applied in the experimental prototype, the documents that are highly relevant to the query but have a low keyword frequency should be ranked ahead. The traditional keyword-based ranking algorithm only considers the frequency of the keyword when evaluating the degree of relevance between the query and the document. After introducing the concept of domain knowledge, the ontology-based ranking algorithm makes it possible to 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.

In this experiment, we use Parkinson's disease as the domain knowledge. The ontology was built by BioPortal [15]. In the Parkinson's disease ontology, it mainly describes the concept of the Parkinson's disease symptoms. For example, the symptoms of Parkinson's disease can be divided into: motor symptom, non-motor symptom and long-term complication of medication. Then, the symptom such as akinesia, tremor etc. are the subclasses of motor symptom.

4.2 Prototype Design

An experimental prototype was built to compare different search approaches. The prototype consists of three parts. First, an index table for the retrieval was built. Second, create a model for the candidate document selection. Third, complete the candidate documents ranking and return them to the users. The workflow of the experimental prototype is shown in Figure 3

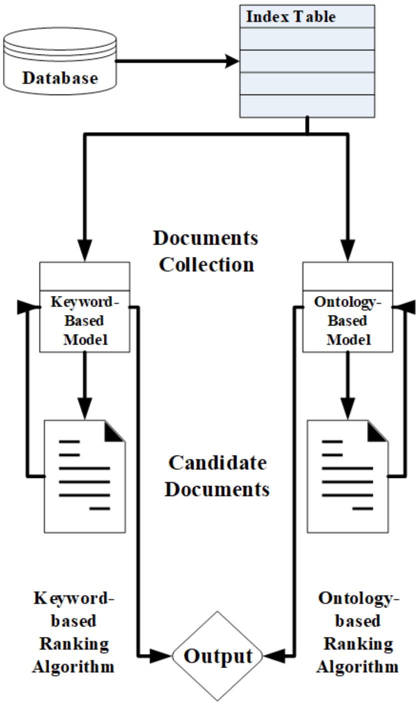


Figure 3. Workflow of the Prototype

DOI: https://doi.org/10.30564/jcsr.v1i2.972

4.2.1 Create an Index Table

When the database is ready (see the detail in 5.1), an index table was set up for every single word on each document in the database.

(1) A program was designed to read every document in the database, and an index table is a 'key-value' pairwise system for retrieving the document in the database by using the index.

(2) Suppose we use path1 to denote the path of the first document in the database, and word1, word2, …, are the single words, they composed the content of the document.

(3) When the program read the word1, the first word in the content of the document. A 'key' was assigned in the index table, which is the word 1. The corresponding 'value' is set to path1, the path of the document in the database. In the prototype, the index table was created by the text files. The filename is index (the 'key) and the content of the text files in the index table is the path of the document. and the same with word 2, word 3, and so on. As the program runs, the program reads every single word in each document in the database, and then assigns the path of the document to the corresponding index. This program will be terminated until every word in each document is processed, the workflow of the index table creation is shown in Figure 4.

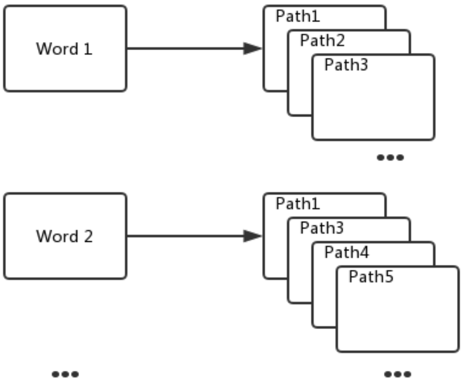


Figure 4. The demonstration of the index table

4.2.2 The Models Comparation

In this research prototype, I used two different models (keyword-based model and ontology-based model) to compare their performance. When the index table created, the models use two different algorithms to generate the candidate documents for the queries.

First, the traditional keyword-based model to select the candidate documents from the index table is based on the query keywords. Namely, when a query is inputted, the traditional keyword-based model extracts the index which is corresponding with the query keyword in the index table, then the model will analyze the document path saved in the extracted index and treat these documents as the candidate documents.

Second, the ontology-based model in this prototype uses domain knowledge to improve the selection of candidate documents and the result's ranking. When a domain knowledge is inputted to the ontology-based model, if the query belongs to the specific domain knowledge, first, the model will analyze the domain knowledge and obtain the relationship between each concept in the knowledge. Then, the candidate documents will consist of all the indexes corresponding to the concepts in the domain knowledge. Which means, the model will not only select the documents under the corresponding query keyword index, but also will selects the indexes that are related to the query keyword (the documents under the relevant concept keyword in the index table). By using this approach, we will obtain the candidate documents that more semantically fit for the query.

Next, research prototype will apply the ranking algorithm to arrange the candidate documents based on the relevance between the candidate documents and query. The candidate documents will get a higher ranking if they are more relevant with the query, and so on. For the traditional keyword-based model, it calculates the relevance of the query and the documents based on the frequency of the keyword appearing in the content of the document. If a keyword appears more frequently in the content of a document, the model will consider that the document is more relevant to the keyword, so the document will get a higher ranking. For the ontology-based model, it calculates the semantic similarity among the concepts in a domain knowledge and add the similarity as a weight in the keyword-based search. Namely, when the model obtains the semantic similarity among every concept in a domain knowledge, the model can know how relevant between every pair of concepts. Then, when the users try to search one of the concepts in the domain knowledge, the model can return to the search engine not only the concept that the user queried, but also its relevant concepts according to the degree of association.

4.3 Query Testing

As introduced in the previous section, a Graphical User Interface of the experimental prototype was implemented (Figure 5). There are four parts in this GUI, wherein the text field above the interface is used to input the que-

DOI: https://doi.org/10.30564/jcsr.v1i2.972

ry keywords. When the query keywords were input by the users, then the users can select either keyword-based search or ontology-based search. These two search buttons correspond to search using keyword-based model or ontology-based model, respectively. Regardless of which search method is used, the results are returned to the text box in the middle of the GUI.

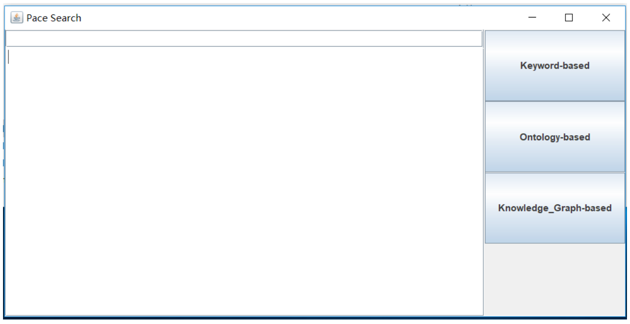
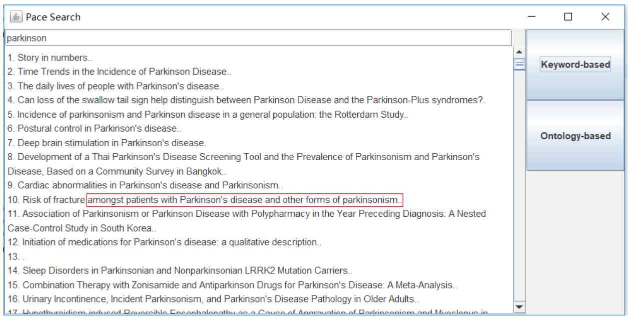


Figure 5. The GUI of experimental prototype

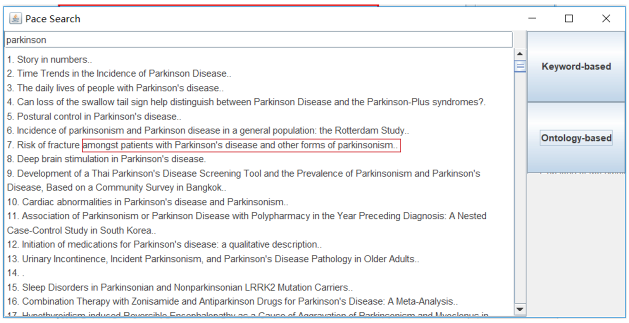
The proposed experiment is used to verify that ontology-based search supported by domain knowledge have better retrieval ability than keyword-based search in a certain field. More specifically, we hope to find out the documents that contain more concepts related to the query keyword via the ontology-based search, and then rank those documents in the top. For example, I used 'Parkinson' as the query keyword and enter it into the experimental prototype GUI. Then I compare the results returned by different search methods. The Figure 6 is the result returned by the keyword-based search, and Figure 7 is the result returned by the ontology-based search. We can observe that the article 'Risk of fracture amongst patients with Parkinson's disease and other forms of parkinsonism', in this article, it includes more concepts related to the query keyword, such as postural instability, falls and so on. If we use keyword-based search, the query keyword appears in the document infrequently, thus the result of the keyword-based search is not ranked first (10th). However, if we use ontology-based search, because of the domain knowledge's support, the ontology-based search mode can retrieve the concept of query keyword in the document. So, when we calculate the ranking score, the model will regard the concepts related to these query keyword as weights, and the documents will be ranked to the front by the ontology-based search model (7th).

By comparing the results of the different model, we confirm the hypothesis two in this chapter, and verifying that the ontology-based model is better than the keyword-based model in performance.



```
parkinson
1. Story in numbers..
2. Time Trends in the Incidence of Parkinson Disease..
3. The daily lives of people with Parkinson's disease..
4. Can loss of the swallow tail sign help distinguish between Parkinson Disease and the Parkinson-Plus syndromes?.
5. Incidence of parkinsonism and Parkinson disease in a general population: the Rotterdam Study..
6. Postural control in Parkinson's disease..
7. Deep brain stimulation in Parkinson's disease.
8. Development of a Thai Parkinson's Disease Screening Tool and the Prevalence of Parkinsonism and Parkinson's
Disease, Based on a Community Survey in Bangkok..
9. Cardiac abnormalities in Parkinson's disease and Parkinsonism..
10. Risk of fracture amongst patients with Parkinson's disease and other forms of parkinsonism..
11. Association of Parkinsonism or Parkinson Disease with Polypharmacy in the Year Preceding Diagnosis: A Nested
Case-Control Study in South Korea..
12. Initiation of medications for Parkinson's disease: a qualitative description..
13. .
14. Sleep Disorders in Parkinsonian and Nonparkinsonian LRRK2 Mutation Carriers..
15. Combination Therapy with Zonisamide and Antiparkinson Drugs for Parkinson's Disease: A Meta-Analysis..
16. Urinary Incontinence, Incident Parkinsonism, and Parkinson's Disease Pathology in Older Adults..
17. Hypothyroidism-induced Reversible Encephalopathy as a Cause of Aggravation of Parkinsonism and Myoclonus in
```

Figure 6. The result of Keyword-based model



```
parkinson
1. Story in numbers..
2. Time Trends in the Incidence of Parkinson Disease..
3. The daily lives of people with Parkinson's disease..
4. Can loss of the swallow tail sign help distinguish between Parkinson Disease and the Parkinson-Plus syndromes?.
5. Postural control in Parkinson's disease..
6. Incidence of parkinsonism and Parkinson disease in a general population: the Rotterdam Study..
7. Risk of fracture amongst patients with Parkinson's disease and other forms of parkinsonism..
8. Deep brain stimulation in Parkinson's disease.
9. Development of a Thai Parkinson's Disease Screening Tool and the Prevalence of Parkinsonism and Parkinson's
Disease, Based on a Community Survey in Bangkok..
10. Cardiac abnormalities in Parkinson's disease and Parkinsonism..
11. Association of Parkinsonism or Parkinson Disease with Polypharmacy in the Year Preceding Diagnosis: A Nested
Case-Control Study in South Korea..
12. Initiation of medications for Parkinson's disease: a qualitative description..
13. Urinary Incontinence, Incident Parkinsonism, and Parkinson's Disease Pathology in Older Adults..
14. .
15. Sleep Disorders in Parkinsonian and Nonparkinsonian LRRK2 Mutation Carriers..
16. Combination Therapy with Zonisamide and Antiparkinson Drugs for Parkinson's Disease: A Meta-Analysis..
17. Hypothyroidism-induced Reversible Encephalopathy as a Cause of Aggravation of Parkinsonism and Myoclonus in
```

Figure 7. The result of Ontology-based model

4.4 Model Evaluation

In order to assess the performance of different models, we need a system to evaluate different models. For the page rank model in search engines, the metrics we need to consider are: Precision, Recall and F1 score.

Precision and recall are two metrics that are widely used in the field of information retrieval and statistical classification to assess the quality of results. The precision is the ratio of the number of related documents retrieved to the total number of documents retrieved. The recall refers to the ratio of the number of related documents retrieved and the number of related documents in the document library. In general, precision is how many of the retrieved items (such as documents, web pages, etc.) are accurate. Recall is how many accurate entries are retrieved. For example, when a search engine returns 30 pages only 20 of which were relevant while failing to return 40 additional relevant pages, its precision is 20/30 = 2/3 while its recall is 20/60 = 1/3. So, in this case, precision is "how useful the search results are", and recall is "how complete the results are".

We can evaluate the web page ranking result by four categories, the following table shows the four categories of the two dimensions.

	Relevant Group	Non-relevant Group
Retrieved	TP (True Positive)	FP (False Positive)
Non-retrieved	FN (False Negative)	TN (True Negative)

DOI: https://doi.org/10.30564/jcsr.v1i2.972

The precision can be expressed as:

$$P(precision) = \frac{TP}{TP+FP} \qquad (7)$$

The recall can be expressed as:

$$R(recall) = \frac{TP}{TP+FN} \qquad (8)$$

For example, suppose we have 60 positive samples and 40 negative samples. We need to find all positive samples. Then, the system retrieved 50 samples, and 40 of them are true positive samples. The indicators we listed above are calculated as follows:

TP: 40, FN: 20, FP: 10, TN: 30

$$P(precision) = \frac{TP}{TP+FP} = \frac{40}{40+10} = 80\%$$

$$R(recall) = \frac{TP}{TP+FN} = \frac{40}{40+20} = 66.67\%$$

4.5 Experiment Result

The keyword-based model retrieved 9569 documents from the database, and 8367 documents are related to the query keyword (Parkinson). The ontology-based model retrieved 9927 web pages, and 8448 of them are related to the query keyword (Parkinson). By the comparison, we can find that the ontology-based model can retrieve more related documents (recall 94.62%). Through the experiments, the hypothesis one claimed in this chapter is confirmed, and we can verify that the proposed ontology-based search model supported by domain knowledge can achieve our expected results.

Conclusion

In this paper, we proposed an ontology-based web pages ranking model. This model combined the TF-IDF weights with the domain knowledge base and introduces the semantic relevance of the keywords and another vocabulary in the web pages. By using the model, we can improve the users' search experience by letting the search engine achieve the capabilities to capture the conceptualizations involved in users' intention and web pages' content meaning. In the future, we will enhance the representation of the knowledge, because the ontology has limited ability to represent the custom relation. We will perfect our model by introducing the knowledge graph, with the goal to achieve more reasonable and accurate web pages' ranking result.

References

[1] L. P. Sergey Brin. The Anatomy of a Large-Scale Hypertextual Web Search Engine [J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.

[2] J. M. Kleinberg. Hubs, authorities, and communities [J]. ACM Computing Surveys, 1999, 31(4).

[3] J. H. O. L. Tim Berners-Lee. The Semantic Web [J]. Scientific American, 2011, 284: 28-37, 17.

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

[5] S. B. a. L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine [J]. Computer networks and ISDN systems, 1998, 30: 107-17.

[6] D. L. McGuinness. Ontologies Come of Age [J]. Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, J. A. H. H. L. W. W. T. B. Dieter Fensel, Ed., MIT Press, 2002.

[7] M.-K. K. B.-L. D. Abdelkrim Bouramoul. An ontology-based approach for semantics ranking of the web search engines results [J]. in International Conference on Multimedia Computing and Systems, Tangier, 2012.

[8] J. R. S. N. Prerna Parmeshwaran. The Use of Ontology in Semantic Search Techniques [J]. International Journal of Computer Applications, 2015, 127(6): 21-24.

[9] A. S. Butt. Ontology Search: Finding the Right Ontologies on the Web [J]. the 24th International Conference on World Wide Web, Florence, 2015.

[10] T. R. Gruber. A translation approach to portable ontology specifications [J]. Knowledge Acquisition - Special issue: Current issues in knowledge modeling, 1993, 5(2): 199-220.

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

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

[13] B. G. S. L. C. B. Joeran Beel. Research-paper recommender systems: a literature survey [J]. International Journal on Digital Libraries, 2016, 17(4): 305-338.

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

[15] 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 [j]. Nucleic Acids Research, 2011, 39: W541-545.

DOI: https://doi.org/10.30564/jcsr.v1i2.972