Global Scientific JOURNALS

GSJ: Volume 8, Issue 1, January 2020, Online: ISSN 2320-9186

www.globalscientificjournal.com

# Semantic Information Retrieval Based on Adaptive Learning Ontology

Waddah Munassar, Amal Fouad Ali

Department of Information Technology

Faculty of Engineering, University of Aden

## Abstract

Information retrieval ranking document is order the documents according to the users searching query. Term frequency (tf) that appears in the document is one of the most existing appoint for information retrieval. Although the term frequency, most of query search is give the result according to keyword search not by semantic search, ranking document may give irrelevant page to the users. Even though the number of times that the term occurrence is more relevant, but not implied for rank documents according to their proximity to learners query. This paper presented a semantic ranking and query that according to the learners profile preferences. The obtained results depicts that the result of learner query is relevant to the learners preferences.

**Keywords**: Ontology; Semantic Ranking; Cosine Similarity; Vector Space Model;



## 1.Introduction

With the increasing amount of documents available online, it is difficult for the users to obtain the required information. A good Information Retrieval system is not only to get the relevant resource for the learners needs, is also for reducing the number of retrieved hits.

One of the most significance process in information retrieval is document ranking algorithm, is used to obtain high efficiency search results. The obtained document ranked according to the highest similarity score of the relevant user query. Term Frequency Inverse Document Frequency approach (TF-IDF) algorithm [1], is an easiest ranking functions and used for weighting a keyword in document. TF-IDF assign the importance to keyword based on the number of times appear in the document. Traditional ranking method for similarity measure is based on vector space model, such as Cosine coefficient, Dice coefficient and Jaccard coefficient.

The limitation of document ranking (keyword-based search) is not enable the search engine to understand the meaning of keyword and differentiate between relevant and irrelevant keywords that appropriate to user's query. Although the term frequency (tf) is compute the term frequency in the document, but not meant rank documents according to user‘s query. To solve the limitations of keyword-based search, semantic search is used semantic similarity measuring through words, concepts or ontologies and became methods to understanding the meaning of keyword. The rest of this paper is organized as follows. Section 2 describes the literature review . Section 3 illustrates an overview of Adaptive Learning Ontology architecture. In section4, discuses Semantic Ranking Approach. Section 5 evaluates the retrieval efficiency of ontology. Conclusion are covered in section6.

## 2. Literature Review

### 2.1 Ontology

Ontology is a formal explicit specification of a conceptualization. Ontologies are formal models that describe a specific domain and determine the meanings of terms by describing their relationships with other terms in the ontology[2]. It is utilized to reason about the properties of that domain and might be utilized to describe the domain. Ontology gives a shared vocabulary, which can be utilized to model a domain that is, the kind of objects, and/ or existing concepts , and their relations and properties. Ontology is used to share common understanding of the structure of information among people or software agents, and to enable reuse of domain knowledge [3].

### 2.2 Term Weighting

Term weighting (TW) is a procedure to compute a weight for each term perform a specific document. This weight reflects to what degree does this term perform that document. Because of its significance, term weighting is utilized in numerous fields such as information retrieval (IR), document clustering, and some more. Term weighting enhances the precision, recall measures and rank of the retrieved documents[4].

There are various of term weighting algorithms, the most popular algorithm is the Term Frequency Inverse Document Frequency ( TF-IDF ) which is a statistical based method which calculate the weight of a term i in document j ( wij ) as illustrate in equation (1).[ 5]

$$w_{i,j} = tf_{i,j} \times idf_j = tf_{i,j} \times \log N/df_j \qquad (1)$$

Where, tfij is term frequency Ti in document Dj , N is the total number of documents, dfj is number of document contain term Ti.

### 2.3 Cosine Similarity

Cosine similarity is a common similarity measure between two vectors of an inner product space that measures the cosine angle between them . The value of cosine 0° is 1, and it is less than 1 for any other angle, so in case of cosine similarity is 1, that means the two documents are entirely similar. On the other hand, documents are dissimilar when the cosine similarity is -1. The calculation of cosine similarity performed by the following formula [6]:

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \qquad (2)$$

where wij and wik are components of vector dj and dk respectively.

### 2.4 Semantic Similarity

Semantic similarity is a method to measure the semantic similarity between concepts or the semantic distance between two concepts according to a given ontology. It enables services to be picked and categorized according to their relate to a given query, and a user's profile and preferences [7].In the last few decades various methods of determining semantic measures have been proposed. The ontology taxonomic hierarchy can be specified with three factors associated, which is : The depth factor, path length factor and local density factor in the hierarchy do affecting the semantic distance measure[7]. one of the Path Based Similarity Measures is Wu & Palmer Measure that is based on the path length between concepts located in a taxonomy, the concepts with greater depth would be more similar (because of specificity). This measure is given by:

$$Sim_{wup}(C1, C2) = \frac{2*N}{N1+N2+2*N} \qquad (3)$$

Where C1 and C2 are concepts in the taxonomy, N1 and N2 are the distance (number of IS-A links), N is the number of IS-A links from C to the root of ontology.

## 3. Adaptive Learning Ontology

This section describes taxonomic hierarchy for Adaptive Learning Ontology which contains ( Learner Profile Ontology Representation and Learning Resource Ontology Representation) and ontology indexing weight.

## 3.1 Ontology Representation

The reason for building ontology is To share common understanding of the structure of information among people or software agents E.g. for communication among sites in ecommerce, to enable reuse of domain knowledge, and to make domain assumptions explicit to avoiding hardwiring into code, and can be changed without changing code. The relationship between ontology concepts make the machine understand the meaning of word not only for readable, that makes ontology used in rank document because the semantic search give all the relevant document for user's query search. This paper is used Adaptive Learning Ontology [8] to retrieve the learning resource according to the learner's style and knowledge and ranked the resources according to the learner preferences.

### 3.1.1 Learner Profile Ontology

Learner profile contains information about learner's personal information, prior knowledge, and learning styles as illustrate in Figure 1. The ontology is defined as classes, namely the learner class which is related to the learning style and knowledge level class through the *belong_to_style, has knowledge* properties as an object property. The class learner is defined *name, birth date, phonNo* and *study-year* properties as Data type property. The learning style class is divided into four subclasses :1. active-reflective class: have two subclasses active and reflective class , 2.visual-verbal class: have two subclasses visual and verbal class, 3. sensing-intuitive class: have two subclasses sensing and intuitive class, 4.sequential-global class: have two subclasses sequential and global class. The knowledge level class has three subclasses beginner , medium and advance class.

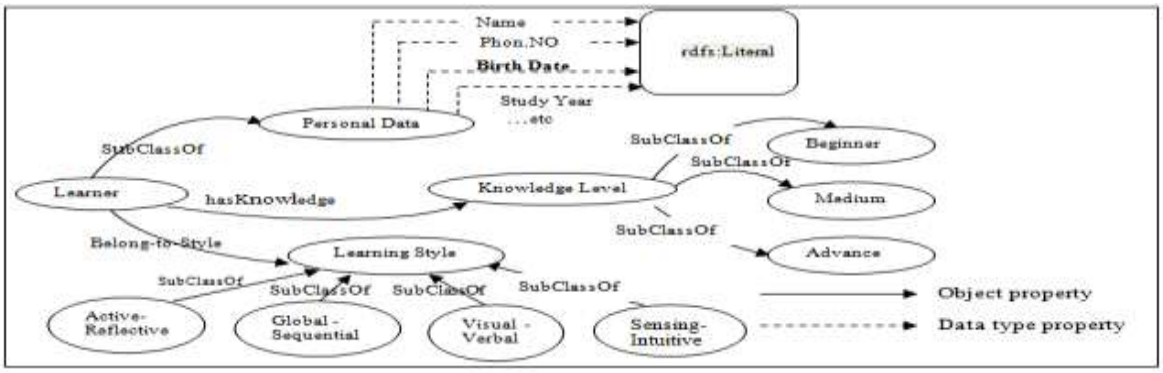


Figure 1. Learner Profile Ontology

### 3.1.2 Learning Resource Ontology

It contains all the knowledge for a particular course, which have many concepts and these concepts can be represented in a form of learning object such as presentations, questions activities, examples, exercises,...etc. The learning resource ontology is illustrated in Figure 2. class learner has *takes* object property used to list the courses taken by the learner and to join between learner and course class. The concept class contains several objects properties like:1- *ccBelongsto* : relate the

concept to its related course, 2-*consistOf* : relate the concept and its sub-concepts, 3- *similerto*: to map between concepts which have same semantic meaning, 4- *oppositeOf* :to map between concepts which have opposite semantic meaning, 5- *nextConcept*: is the next concept possible to the given concept, 6- *previousConcept* : the previous concept of the current concept, 7- *hasrequisite*: the concepts may to know before start study concept, 8- *isprerequisiteFor* :it is inverse of *hasrequisite* and denote the concepts for which it is a prerequisite for, 9- *isdescribedby* explain a concept by using digital resources and it is opposite of *describe* property in the resource class. The *conceptname* is a data type property for class concept to define the concept name. The resource class has objects properties like : 1- *support* : is to relate the resources to the learning style, 2- *suggest*: is to suggest the resource of learning object to the learner according his style, 3-*ProvideTo*: provides the resource of learning Object to the learner according to knowledge level of him, 4- *Includedin*: it is resources included in a course and it is inverse *of hasResource* , 5- *describes*: has inverse relation with *isdescribedby*, this property relate the resources to the concepts ,6- *hasDescription*: it is to join between the resource and its descriptions. The course class has objects properties like:1- *hasconcept*: which joined the course and its related concepts, it is also has inverse relation of *ccBelongTo*, 2- *hasresource* :denote to the set of resources which compose a course. The *courseName* and *courseDescription*, are a data type property. The Resource Description class has two objects properties: 1- *difficultlevel* property is for determine the knowledge level of resource, 2-*helptoachieveknowledge* property is for join resource description with the concepts. The Resource Description class is also contain of data properties such as *createdby, hasKeyword, type*, and *language*.

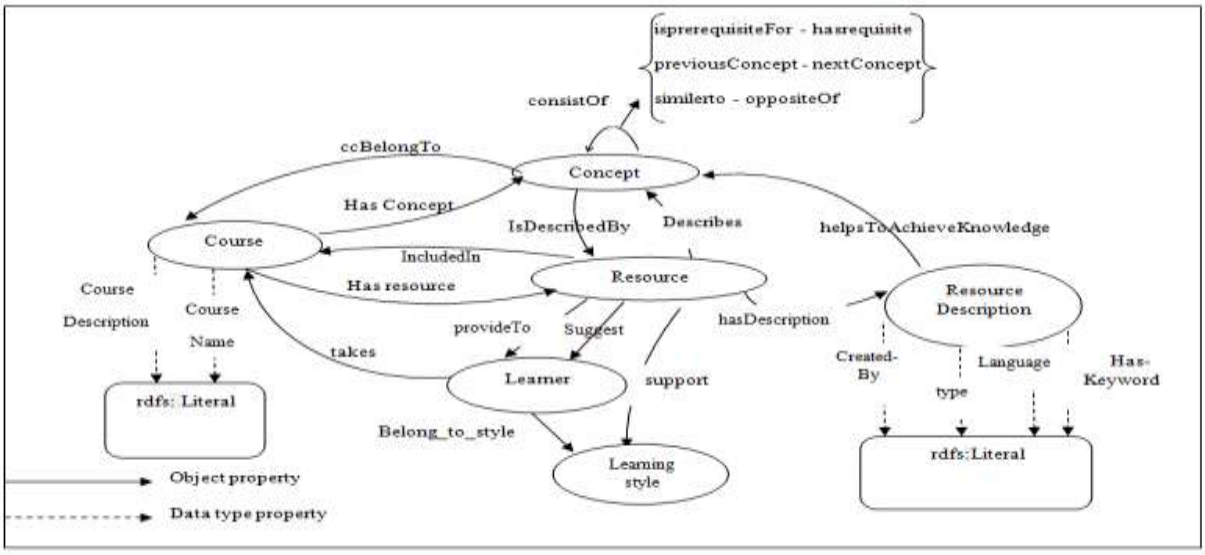


Figure 2. Learning Resource Ontology

## 3.2 Ontology Indexing

The hierarchy weight of subsumption (hypernym/ hyponym or meronym/holonym hierarchy) in Adaptive Learning ontology is measured based on Wu and Palmer measure .These weights are shown in Table 1. The weight was assigned as 1 if query keyword and ontology keyword were the same word or synonymous and assigned as 0.8 if query keyword and document keyword have same sub area (ex. JavaSE Self-assessment and Java EE Self-assessment). The weight was assigned as 0.6 query keyword and document keyword have same area (ex. JavaSE Self-assessment and Java SE multiple-question )and assigned as 0.4 if query keyword and document keyword have different area (ex. JavaSE Self-assessment and Java pages). The weight is assigned to 0 if the query keyword not found in ontology. Adaptive Learning ontology Weight illustrated in Figure 3.

Table 1. Adaptive Learning ontology Weight based on Wu and Palmer Measure

| Relationship Type | Weights |
|---|---|
| **Repetition /Synonymy** | **1** |

| | |
|---|---|
| **Same sub area** | **0.8** |
| **Same area** | **0.6** |
| **Term or Keyword on adaptive Ontology** | **0.4** |
| **not found In adaptive ontology** | **0** |

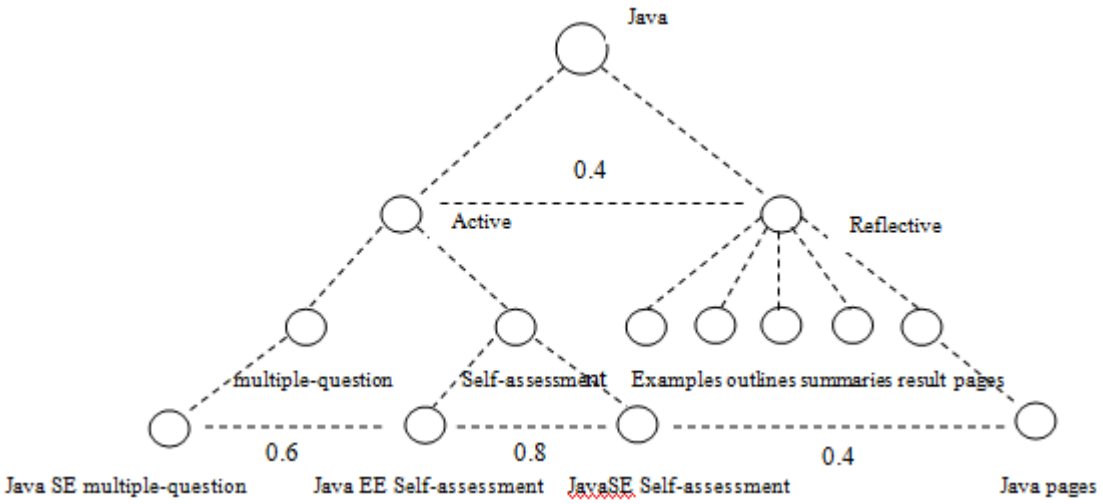


Figure 3. Adaptive Learning ontology Weight

## 4.Semantic Ranking Approach

Semantic ranking describes the document semantic score to rank documents according to query-document matching scores that use tf-weight and adaptive learning ontology weight. The following steps describe the process of document semantic ranking[1].

• The document as vector process is a weight of term frequency that computed by formula (1).
• The query as vector process is a weight of term between query-document that using adaptive learning ontology weight (table 1).
• The document similarity computation between the query and document vector by Cosine similarity measure (3) that is a measure of similarity between two vectors of an inner
product space.
• Final similarity score between query and document by formula (4).

$$Simscore\,(d) = w_{t,d} * w_{t,q} \qquad (4)$$

## 4.1 Rank and Query Processing and Searching

The query language of ontology is SPARQL protocol. The search procedure starts with the user's keyword query. The keyword is transform to formal SPARQL query, which returns a list of instance tuples that satisfy the query. The query result in Figure 4 shows the query in Apache Jena Fuseki server which is using for SPARQL endpoint and triple store, that obtained the learning resource to the learner query and rank the result according to the weight of each resource.

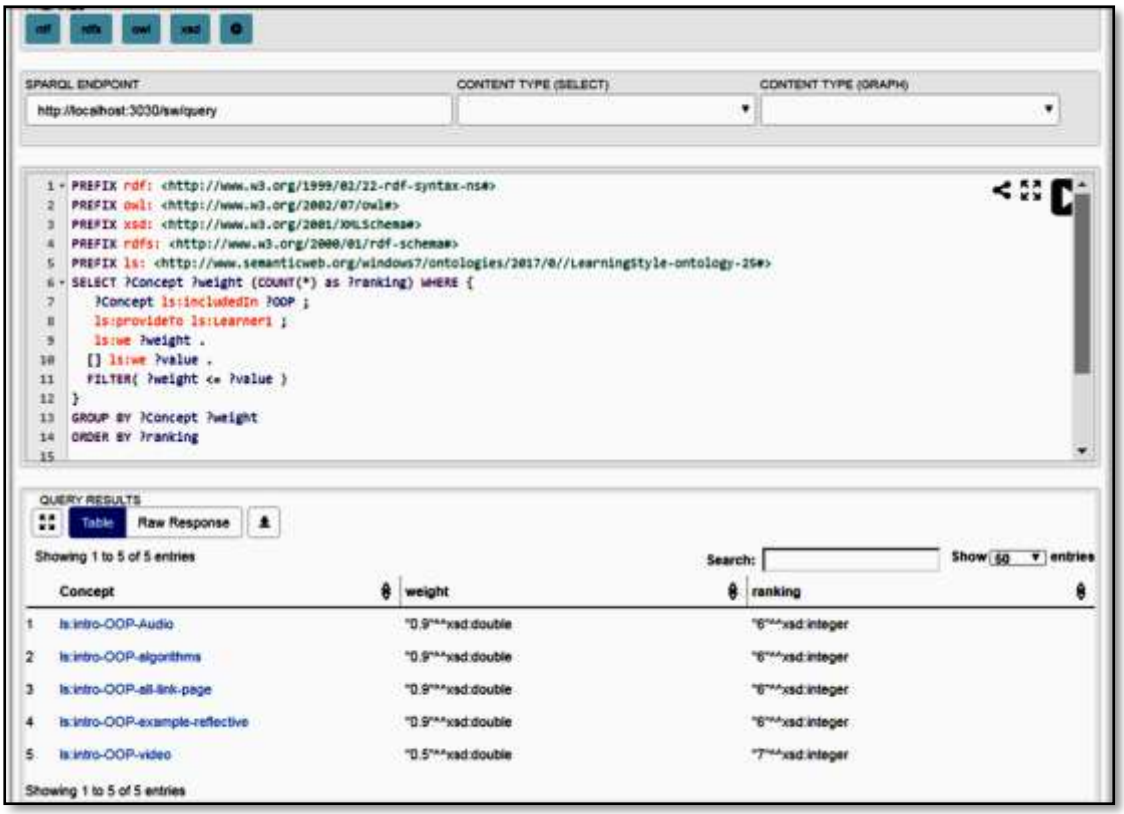


Figure 4.Ranking Sparql Query Result

## 5.Evaluation

The evaluation information retrieval is the rate of success of receipt the learning resource to the learner according to learning styles and knowledge level as shown in Table 2. The precision and recall are using to measure the evaluation in order to determine the retrieval efficiency. Precision can be define as the number of learning resources retrieved that are relevant to the learning styles of the learner [9]. Recall can be define as the number of relevant of learning resources and are successful retrieved. The F-measure is efficient overall representation of precision and recall, as shown in Table 3. The Precision and Recall can be calculated as :

Precision = A/A+c (5) ,

Recall = A/A+B (6).

(A) is denote as the number of retrieved resources that are relevant, (B) is the number of relevant resources that are not retrieved and C is the number of retrieved resources that are not relevant. The F-measure is obtained by using Precision and Recall : F-measure = 2[ (Precision * Recall) / (Precision + Recall) ]. The evaluation result in Table 3 means that the retrieved learning resource Has strong number of the chosen a relevant resources for the learner.

Table 2: Evaluation results of the experiment queries

| Query | No of Relevant Resources | Total No of results shown |
|---|---|---|
| Q1 | 25 | 29 |
| Q2 | 34 | 34 |
| Q3 | 20 | 20 |
| Q4 | 27 | 29 |
| Q5 | 86 | 86 |
| Q6 | 20 | 20 |
| Q7 | 20 | 20 |
| Q8 | 30 | 35 |
| Q9 | 49 | 50 |
| Q10 | 86 | 87 |

The evaluation result in Table 3 is compared between the current result and previous result [10], the average value of precision and recall in [10] is 0.72 and 0.49 and in this study the average value of precision and recall is 0.95 and 1, that's means the retrieved learning resource in adaptive learning ontology has strong number of the chosen a relevant resources to the learner.

Table 3: Evaluation Result

| Query | Semantic Search | | Semantic Search | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| Q1 | 0.86 | 1 | 1 | 0.50 |
| Q2 | 1 | 1 | 0.80 | 0.65 |
| Q3 | 1 | 1 | 0.75 | 0.50 |
| Q4 | 0.93 | 1 | 0.55 | 0.45 |
| Q5 | 1 | 1 | 0.50 | 0.35 |
| Average | 0.95 | 1 | 0.72 | 0.49 |

- **Evaluate By ROC**

The query results been evaluated by using Receiver Operating Characteristic (ROC) to measure the quality of the query results the representation of query result is shown in Table 4. ROC is a graphical curve plotted using the True Positive Rate (TPR)(7) and the False Positive Rate (FPR)(8) of the classification results. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 − specificity).

The True Positive Rate and False Positive Rate values have been calculated and presented in Table 5.

$$TPR = \frac{TP}{(TP + FN)} \qquad (7)$$

$$FPR = \frac{FP}{(FP + TN)} \qquad (8)$$

Table 4: Representation of Query Result

| Cut Scores: | Most Normal ← 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 → Most Abnormal | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Normal Cases | 20 | 20 | 20 | 34 | 86 | 86 | 49 | 27 | 25 | 30 | 397 |
| Nonnormal Cases | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2 | 4 | 5 | 13 |
| | 20 | 20 | 20 | 34 | 86 | 87 | 50 | 29 | 29 | 35 | 410 |

Table 5: True Positive Rate and False Positive Rate values for ten Queries.

| True Positive | False Positive | Area |
|---|---|---|
| 0.0504 | 0.0000 | 0.0000 |
| 0.1008 | 0.0000 | 0.0000 |
| 0.1511 | 0.0000 | 0.0000 |
| 0.2368 | 0.0000 | 0.0000 |
| 0.4534 | 0.0000 | 0.0432 |
| 0.6700 | 0.0769 | 0.0563 |
| 0.7935 | 0.1538 | 0.1274 |
| 0.8615 | 0.3077 | 0.2748 |
| 0.9244 | 0.6154 | 0.3701 |
| 1.0000 | 1.0000 | - |
| AUC → | | 0.8716 |

ROC curves representing the query results are shown in Figure 5. The area under curve AUC is 0.8716 that's means the framework is considered good as appeared in Table 6 [11], to retrieve the relevant learning object to the learner.

Table 6: Categorization of ROC Curves.

| AUROC | Category |
|---|---|
| 0.9-1.0 | Very good |
| 0.8-0.9 | Good |
| 0.7-0.8 | Fair |
| 0.6-0.7 | Poor |
| 0.5-0.6 | Fail |

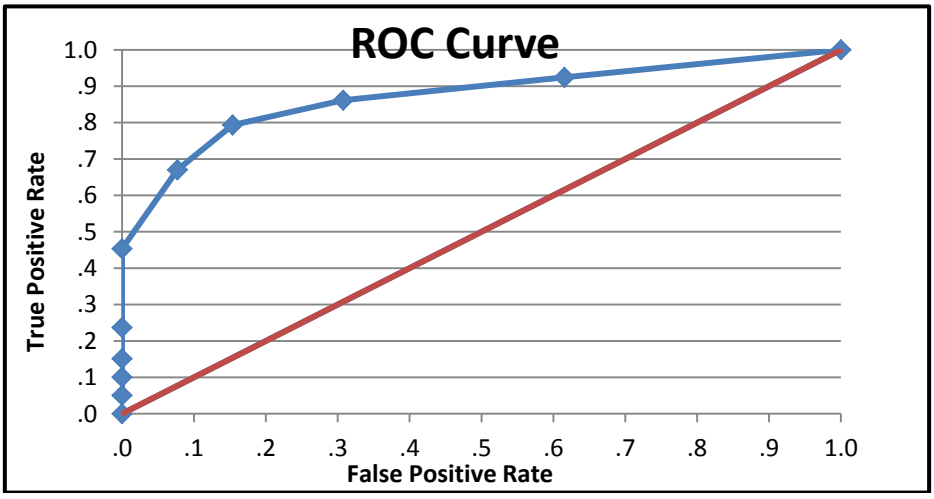


Figure 5. ROC Curve for The Query Result

## 6. Conclusion

The presented study here has successfully met its objective to overcome the key word based search limitations by depends the work on the ontological modeling in the information retrieval**.** The semantic web improves the information retrieval quality by returns more accurate and related results to satisfy the user needs (completeness and accuracy) . In brief, the main achievements of this work is to proposed an ontology based information retrieval approach to enhance the information retrieval performance accuracy as shown above. Secondly, we have proposed to use the ranking algorithm (TF-IDF) for better recall and precision and rank the query according to the user profile and preferences.

## 7. References

1. Thanyaporn Boonyoung, Anirach Mingkhwan." Semantic Ranking based on Computer Science Ontology Weight ".Conference Paper ·IEEE August 2014**.**

2. R. Deepa, Dr. R Manicka Chezian, " An Ontological Approach for the Semantic Web Search and the Keyword Similarity Metrics ". International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

3. Diana Man, " ONTOLOGIES IN COMPUTER SCIENCE ". DIDACTICA MATHEMATICA, Vol. 31(2013), No 1, pp. 43-46.

4. Zeinab E. Attia." A Fuzzy Ontology-based Term Weighting Algorithm for Research Papers".Advances in Information Science and Applications - Volume I. ISBN: 978-1-61804-236-1.

5. DIK L. LEE." Document Ranking and the Vector-Space Model". 0740-7459/97/ 1997 IEEE.

6. Alfirna Rizqi Lahitani, Adhistya Erna Permanasari, Noor Akhmad Setiawan." Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment ".2016 IEEE.

7.Thabet Slimani." Description and Evaluation of Semantic Similarity Measures Approaches". International Journal of Computer Applications (0975 – 8887) Volume 80 – No.10, October 2013.

8. Waddah Munassar, Amal Fouad Ali." Semantic Web Technology and Ontology for E-Learning Environment". Egyptian Computer Science Journal Vol. 43 No.2 May 2019 ISSN-1110-2586.

9. Saowaluk Thaiklang, Ngamnij Arch-Int, Somjit Arch-Int." Learning Resource Recommendation Framework Using Rule-Based Reasoning. Approach ". Journal of Theoretical and Applied Information Technology, Vol. 69 No.1.10th. 2014.

10. Patsakorn Singto." Semantic Searching IT Careers Concepts Based on Ontology ". Journal of Advanced Management Science, Vol. 1, No. 1, March 2013

11. Patricia E. Garrett, Fred D. Lasky, Kristen L. Meier, User Protocol for Evaluation of Qualitative Test Performance; Approved Guideline 2nd Edition. Vol. 28 No. 3. 2008.