Semantic Information Retrieval: An Ontology and RDF-based Model

S. Mahaboob Hussain
Assistant Professor, CSE
Vishnu Institute of Technology
Bhimavaram, India

Prathyusha Kanakam
Assistant Professor, CSE
Vishnu Institute of Technology
Bhimavaram, India

D. Suryanarayana
Professor, CSE
Vishnu Institute of Technology
Bhimavaram, India

Swathi Gunnam
PG Scholar
Vishnu Institute of Technology
Bhimavaram, India

Sharmela S.
PG Scholar
Vishnu Institute of Technology
Bhimavaram, India

ABSTRACT
Retrieving the specific knowledge from the Web becomes a challenging task as it contributes enormous amounts of unorganized textual data. This paper focuses on lessening the time consumed by the user for searching the documents and providing the results as per user intention. This paper demonstrates a semantic query language SPARQL to extract the data from the student career knowledge base constructed using Resource Description Framework (RDF) that gives relevant information to the user. This paper provides the requested information by understanding user's query intention with the created career ontology using RDF and SPARQL in a semantic manner.

General Terms
Semantic Information Retrieval, Semantic Web, Cognitive analysis, Machine Learning

Keywords
RDF, SPARQL, Semantic Web, Semantic Search, Information Retrieval

1. INTRODUCTION
Nowadays, gathering of useful information from the Web is becoming a challenging issue for users. The first generation of the internet in late 1980's was so called the era of computer processing. At that time to get information user has to open a terminal and login to a remote system then look for the file on remote system when it is found it has to be downloaded on local system from the remote system for the reading purpose which is bit complicated. Disadvantages with this process are information access requires expert knowledge about commands, special languages and syntaxes and it is more expensive. To overcome the disadvantages in the first generation of the internet in 1990's the worldwide Web had introduced. To get information user has to open a browser which provides a graphical user interface to the user where the user has to type address or details about the document that document contains hyperlinks which connect documents with each other by clicking on links document is obtained that contains data which is connected to the previous document.

Information navigated from the Web server in this way without expert's knowledge. Information retrieval becomes usable because of search engines. This is much more convenient than the previous one but there is a rich growth in Web this has a creating problem for finding the data on the Web. To overcome this problem, a step towards the semantic Web is required, which is the extension of current Web [1]. Until now, traditional search engines provide approximate results but not exact results in some cases. The Web provides solutions to the user by matching the keywords in a query to the documents available on the internet. The basic concept of this project is to provide the exact information to the users' requested query.

In this paper, authors provided a career pathfinder, which give suggestion to students who has completed tenth, inter and degree. It includes jobs and courses available after completion of their studies. Ontology had designed to create this career path because ontology limits complexity and it will organize information. Ontology provides the abstract model of the domain [2]. At first, authors have studied and done research on the semantic Web and its technologies like RDF, OWL, SPARQL, and the sage of these technologies, various advantages of these technologies. After that information about the different courses for students and details about their entrance exams, jobs, course duration and their related exams etc. have collected.

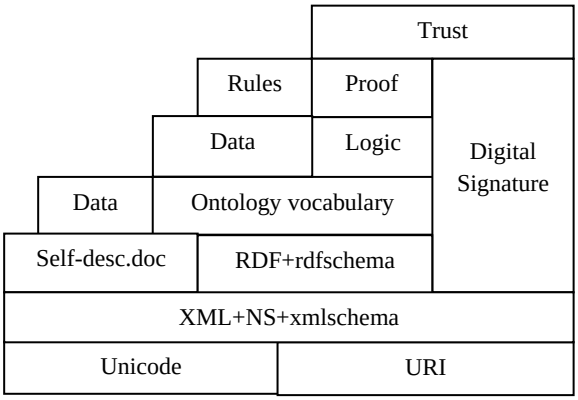


Fig 1: Semantic Web 3.0 architecture

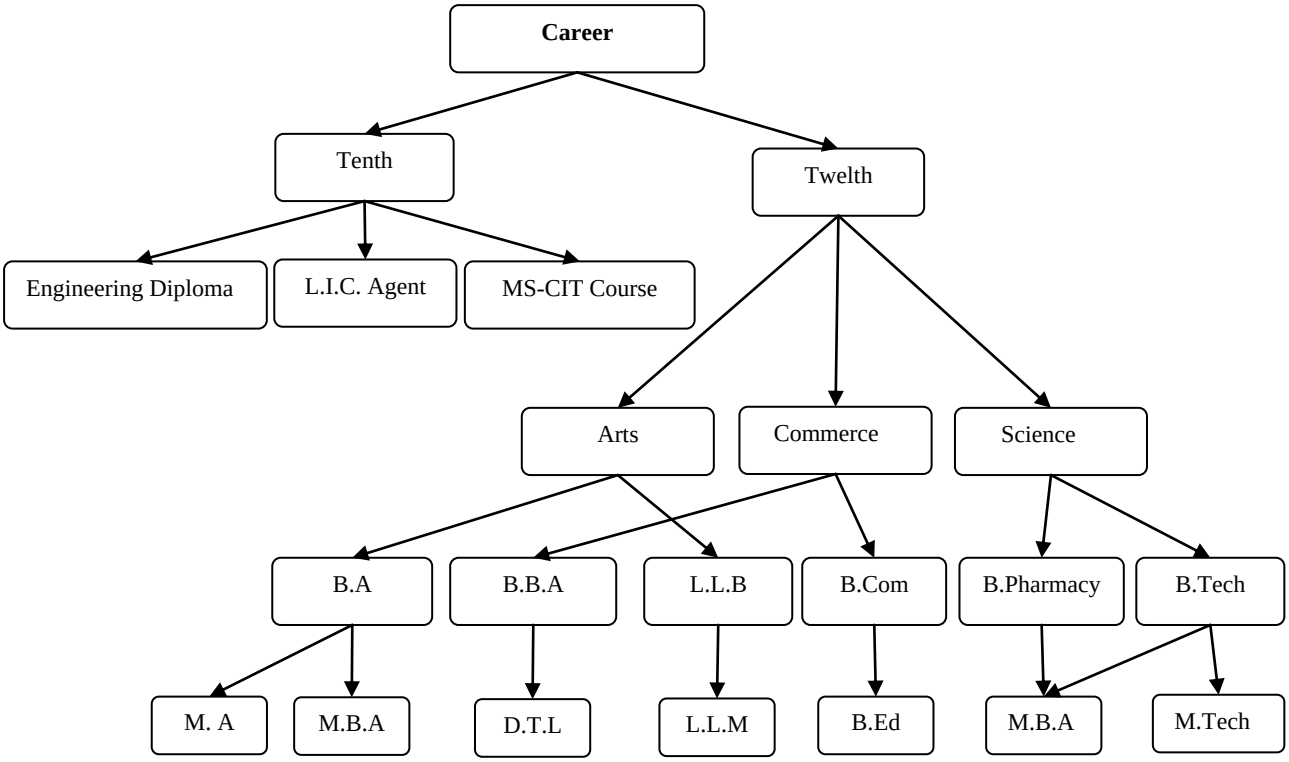


Fig 2: Knowledge representation for education ontology

Semantic Web states that, it is not a separate Web it is an extension of the Web2.0, in which information that given in a completely defined meaning, better enabling computers and people to work in cooperation. The semantic Web is a Web of knowledge connections as in Figure 1. The Web of data is an upgrade of a Web of documents with a huge decentralized database of machine accessible data.

2. DESIGN

2.1 Ontology

Ontology is a meaning of all concepts defined in a machine understandable specification of a shared abstract model. Ontologies represented via classes, relations and instances. Ontologies used to share a common understanding of people and machines. Explicit means the meaning of all concepts that must defined. Ontology's can represent via classes, relations and instances. There are different levels of symbolic forms for describing ontologies from informal range to formal range. There are different types of ontologies according to the level of generality.

- Top level, upper or foundation level ontology represents general concepts and anything in the world can be classified
- Domain ontology focuses on a specific domain which is defined in an upper ontology are described more specifically
- In task ontology, fundamental concepts like task and general activity are described
- Application level ontology focuses on specific task and domain

2.2 Resource Description Framework

Knowledge representation is more expressively using RDF schemas. RDF schema also called as RDF vocabulary description language and used to describe vocabularies for the models that are dealing with [3]. RDF schema allows the definition of classes via *rdfs:Class* and class instantiation in RDF via *rdf:type*. For example, consider the following syntax.

:Career rdf:type rdfs:class .

:Tenth rdf:type :Career.

:Twelth rdf:type :Career.

Here, career path (Figure 2) is an rdf type *class* and Tenth, Twelth are the resources which have type *Career*. Thus, Tenth and Twelth are the instances or members of a class. Besides classes, properties of classes also defined and the properties connect with either classes or literals. *rdf:Property* is used to define a property. On property, restrictions on domain and range can be defined according to type via *rdfs:domain* and *rdfs:range* in rdf schema. In these resources along with hierarchical relationships like subclasses, superclasses along with sub-properties and super properties defined in [4].

3. EXPERIMENTAL ANALYSIS

Definition of properties via *rdf:property* and property restrictions on domain and range via *rdfs:domain* and *rdfs:range* are declared. For example, consider the following syntax.

:Tenth rdf:type rdfs:class.

:courses after rdf type rdf:property.

:courses after rdfs:domain :Tenth .

:courses after rdfs:range :Twelth .

Hierarchical relationships are defined as, *Subclasses and superclasses via rdfs:subclass of.* For instance, consider the following.

:Tenth rdfs:subclass of:Career.

Subproperties and superproperties via subPropertyof

Example:engineeringdiploma rdfs:subpropertyof:Courses

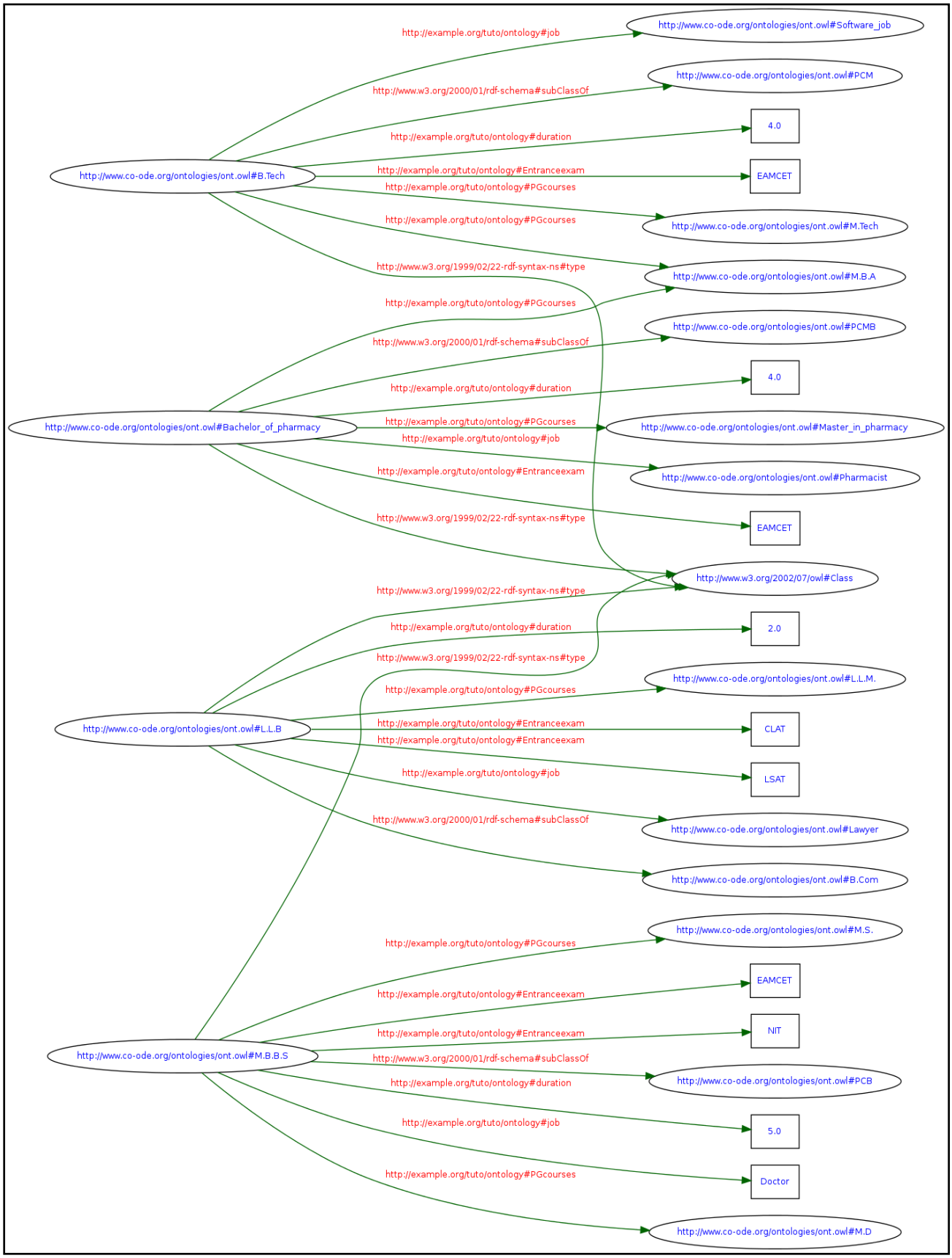


Fig 3: Ontology of RDF relationships among Subject, Object and predicate with URI representations

Constructing the ontology using the RDF is an optimal solution. Then, the semantic information retrieval is easily. The construction of RDF is as below.

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF [
   <!ENTITY dbp "http://dbpedia.org/property/" >
   <!ENTITY dbpedia "http://dbpedia.org/resource/" >
   <!ENTITY dbo "http://dbpedia.org/ontology/" >
   <!ENTITY owl "http://www.w3.org/2002/07/owl#" >
   <!ENTITY tto "http://example.org/tuto/ontology#" >
   <!ENTITY ttr "http://example.org/tuto/resource#" >
   <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#" >
   <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#" >
   <!ENTITY ont "http://www.co-ode.org/ontologies/ont.owl#" >
   <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
]>
 <owl:Class rdf:about="&ont;M.B.B.S">
      <rdfs:subClassOf rdf:resource="&ont;PCB"/>
      <tto:duration
rdf:datatype="&xsd;decimal">5.0</tto:duration>
      <tto:job>Doctor</tto:job>
      <tto:PGcourses rdf:resource="&ont;M.D"/>
      <tto:PGcourses rdf:resource="&ont;M.S."/>
      <tto:Entranceexam>EAMCET</tto:Entranceexam>
      <tto:Entranceexam>NIT</tto:Entranceexam>
 </owl:Class>
    <owl:Class rdf:about="&ont;B.Tech">
      <rdfs:subClassOf rdf:resource="&ont;PCM"/>
      <tto:duration
rdf:datatype="&xsd;decimal">4.0</tto:duration>
      <tto:Entranceexam>EAMCET</tto:Entranceexam>
      <tto:PGcourses rdf:resource="&ont;M.B.A"/>
      <tto:PGcourses rdf:resource="&ont;M.Tech"/>
      <tto:job rdf:resource="&ont;Software_job"/>
   </owl:Class>
```

RDF information is represented by triples and it consists of *subject*, *object* and *predicate*. The RDF graphs are written using RDF/XML. Triples together form RDF Graph. A graph is having triples and literals [5].

An example of a triple is as below.

rdf:subject-the described resource

rdf:predicate-the original resource

rdf:object-the value of the resource

These triplets are represented with URIs from the database as shown in below.

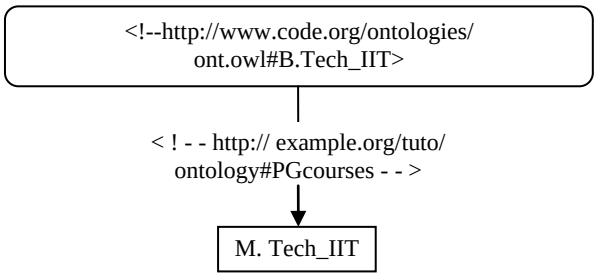
SUBJECT:

<!--http://www.code.org/ontologies/ont.owl#B.Tech_IIT>

PREDICATE:

<!-- http://example.org/tuto/ontology#PGcourses-->

OBJECT: *M.Tech_IIT*



In the above example, RDF triple in the graph representation describing that B.Tech_IIT has PGCourse M.Tech_IIT.

When query applied on the data, it will end as variables, the variables named as character strings, and variable has a question mark at the beginning. SPARQL also constitutes '*select*' statement to extract the data. In addition, retrieved results displayed in a tabular form.

The SPARQL is a query language, which has to traverse through RDF graph as in Figure 3. It uses RDF Turtle serialization and basic graph pattern matching. The graph pattern is triple that contains RDF triple variables that subject or property or object. It is nothing but turtle statement plus RDF variables. Let us look at the example, which is a basic graph pattern, is a set of a triple.

Example: (to look for courses and their jobs)

?courses dbo:jobs ?jobs .

SPARQL QUERY:

```
prefix dbo: <http://dbpedia.org/ontology/>
prefix dbp: <http://dbpedia.org/property/>
prefix dbpedia: <http://dbpedia.org/resource/>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix tto: <http://example.org/tuto/ontology#>
prefix ttr: <http://example.org/tuto/resource#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
select ?twelth?duration?job?courses where {
   ?twelth rdf:type dbo:Twelth .
         ?twelth tto:duration ?duration .
   ?twelth tto:job ?job .
   ?twelth tto:courses ?courses .
   ?twelth tto:EntranceExam ?Entrance Exam
```

Table 1. Representation of the result for SPARQL query

Twelth	Duration	Job	PG Courses	Entrance Exam
B.Tech	4.0	Software job	M.B.A M.Tech	EAMCET
L.L.B	2.0	Lawyer	L.L.M.	CLAT LSAT
B.Pharmacy	4.0	Pharmacist	M. Pharm	EAMCET
M.B.B.S	5.0	Doctor	M.D. M.S.	EAMCET NIT

Authors provide the solutions to the user request by applying SPARQL queries on their data set and the results obtain as in Table1 as SPARQL allows the query consists of triplet format. In addition, the triplet with these *subject*, *object* and *predicate* directly matches with user's request. To define a query, it should come across with many prefixes like RDF to define prefixes and have a keyword prefix to write prefix identifiers. Then, namespaces like *dbpedia*, *rdf*, *dbpedia/ontology* are defined for classes and properties from where they come from [6].

Next part of the query is a select statement, which contains output variables like names of authors and titles then queried on Web, specified by the RDF graph. The query will be imposed on SPARQL endpoint using user interface may have different RDF or dbpedia graphs and many data sets. Then a query is specified using *where* clause within braces. Then, basic triple patterns or graph patterns of the queries obtained to acquire semantic results [7].

To find relevance and analysis of the retrieved results, Palazzo Matrix Method is applied [8] based on these search query logs of the users [9].

4. CONCLUSION

In this paper, a career path presented for the students for who seeks the best career option in future. This career pathfinder includes RDF, Ontology and SPARQL queries. RDF data is used to provide the result for the SPARQL query. All students who had completed their education and stepping towards their further studies and trying to choose their career accordingly can benefit by this career pathfinder. Rather than searching using search engines on the Web, which provides lot of links and the data, user needs to refine the retrieved results and analyse for the actual content that he requires. This career pathfinder application along with the RDF and SPARQL technology, direct semantic results will retrieve. Thus, user can be satisfied with the retrieved results as per their intention. This framework further developed by adding information about the colleges and institutions details that offers admissions and about the available courses.

5. ACKNOWLEDGMENTS

This work has been funded and supported by the Department of Science and Technology (DST), Govt. of India under the Grants No. SRC/CSI/153/2011.

6. REFERENCES

[1] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Sci Am. 284, 34-43 (2001).

[2] Dai, W., You, Y., Wang, W., Sun, Y., Li, T.: Search Engine System Based on Ontology of Technological Resources. JSW. 6, (2011)

[3] Marzano, G.: Using Resource Description Framework (RDF) for Description and Modeling Place Identity. Procedia Computer Science. 77, 135-140 (2015)

[4] Kim, K., Moon, B., Kim, H.: R3F: RDF triple filtering method for efficient SPARQL query processing. World Wide Web. 18, 317-357 (2013)

[5] Jeon, M., Hong, J., Park, Y.: SPARQL Query Processing System over Scalable Triple Data using SparkSQL Framework. Journal of KIISE. 43, 450-459 (2016)

[6] Segaran, T., Evans, C., Taylor, J.: Programming the Semantic Web. O'Reilly, Beijing (2009)

[7] Suryanarayana, D., et al. "Stepping towards a semantic Web search engine for accurate outcomes in favour of user queries: Using RDF and ontology technologies." 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). IEEE, 2015.

[8] Hussain, S. Mahaboob, et al. "Palazzo Matrix Model: An approach to simulate the efficient semantic results in search engines." Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on. IEEE, 2015.

[9] Suryanarayana, D., et al. "Cognitive Analytic Task Based on Search Query Logs for Semantic Identification" IJCTA, 9(21), pp. 273-280, 2016.