International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S, August 2019

Semantic Indexing and Concept Based Methods for Ontology Based Information Retrieval using Dbpedia

P.Priya

Abstract:- ***Web contains immense extent of records. It is helpful for the clients who need to look through data with respect to direction, business, redirection, remedial organizations, topography, and so forth... Looking through such data is essential utilizing particular web records yet looking through obvious data which is required is a horrible errand. This can be developed by semantic pursue which means hugeness or thought based intrigue. In the proposed work, thinking is made utilizing the parts and the properties of the substances. An introspective philosophy based data extraction framework is sorted out which is unequivocal to soccer district. Utilizing the crept data from DBPedia, the perspective is made and the OWL files are populated. Semantic mentioning is then performed on these reports. At long last the mentioning results and the perspective are mapped and the outcomes are appeared to the client dependent on the arranging performed utilizing two segments 1) Frequency of watchword and 2) huge inquisitive words. The introduction of the proposed framework is assessed against the present structures.***

Index terms: Semantic Web; OWL documents; Ontology; DBpedia; semantic mentioning.

I. INTRODUCTION

Web search is the way toward detaching the data on the web. The web record gives the most ideal approach to manage looking at the web for data and its substance makes and changes each day. Looking through data in the web isn't the issue at any rate finding applicable data is especially tricky. In the event that the client does not have any space information where he/she is going to look, by then he/she gives powerfully far reaching solicitation to the web document. The web document accomplishes general page about the client demand and the client need to discover other pursue terms from the site that can assist the client with finding more data on the subject.

In the web, unending number of reports are available. Therefore, scanning for the suitable accounts is a puzzling undertaking. Regardless of the manner in which that many web search gadgets are there to help the reports recovery from the web, clients need to steadily reformulate the solicitation until required records are recovered. More endeavors and time are required to confine the fitting reports. The power practice in data recovery is a catchphrase based solicitation. Such sort of solicitation misses the authentic semantic data in the substance. To decimation such issues, ontologies are made to address information. Legitimately a-days ontologies are the foundation of semantic web applications. There are four classes in semantic tending to 1) catchphrase based 2) structure based 3) see based and 4) customary language based framework [6-7]. The two data extraction and recovery procedure can profit by such information which offers criticalness to the plain message.

Resulting to having gotten the semantic information tended to as hypothesis, the going with stage is tending to. Various solicitation tongues are accessible. Among which SPARQL is before long the top level inquiry language for semantic web. In any case, the clients don't consider such dialects. Thusly, semantic web advances toward streamlining the course toward incorporating the request for the clients.

The remainder of the paper is shaped as looks for after: In zone II, related works are investigated. Section III delineates the proposed framework to vanquish the issue in watchword based solicitation. Part IV portrays about exploratory outcomes. Part V closes the paper.

II. RELATED WORK

One of the most basic issues in catchphrase based way of thinking is the recovery of insignificant information. Henceforth, the criticalness of the client's request must be dissected and the outcomes are to be exhibited appropriately. Here, some essential foundation data is given about the particular key bits of this paper.

A .Traditional Approaches

The old style or standard watchword set up together data recovery approaches depend concerning the IR models. In the zone of substance recovery, no extraction or comment tries are fused and the consequences of the sentences are rejected.

B .Query and Index Expansion Methods

The semantic recovery has the initiation by utilizing WordNet equivalent word sets (synsets) for word semantics. The basic objective is to widen reports and solicitation with the centrality of the words to accomplish better review and precision.

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
P.Priya, Assistant Professor, Department of Computer Science and Engineering BIET, Hyderabad, Telangana, India. (Email: priyaponnusamy29@gmail.com)

C .Semantic Approaches

With the presentation of semantic web impels, information portrayal has wound up being ceaselessly dealt with and present day, which requires also made extraction and recovery techniques to be executed. The other elective approaches for data extractions are structures/rule-based data extractor against overwhelming computational expense. Redone frameworks [8] are favored looked the manual ones considering the exertion spent on the space. The frameworks in [2-4] use hand-made measures to evacuate data. Hand-made measures are also utilized in semantic comment [9]. Wessman et al. [10] depends in a general sense on standard articulations. The general methodology stores the ousted information in RDF or OWL game-plan and SPARQL is utilized to demand that RDF record. Despite the way that this framework offers the better accuracy and overview execution, it is multifaceted solicitation language for the clients.

Along these lines, to vanquish the burdens of learning a formal request language, diverse solicitation interface techniques are proposed. Burst is a recovery framework that can demand semantic information with SPARQL solicitation made from catchphrases. It contains three standard advances: term mapping, demand chart improvement and question arranging. The mapping step attempts to facilitate solicitation terms to ontological assets.

Commonly, more than one solicitation is built by virtue of the ambiguities. In this way in the last advance, the request are assessed with a probabilistic arranging model. The SPARQL delivers try to vanquish any block between catchphrase solicitation and formal solicitation.

D .Semantic Indexing Methods

The elective technique to address headway from catchphrases is semantic mentioning in which the watchword solicitation are immediate encompassed by utilizing the semantic information in RDF learning puts together which are with respect to an extremely essential level recorded. This methodology is finished in OWLIR [16]. It includes a data extraction module, an inferring module and a recovery module. The assessment displayed in [1] can be reached different spaces likewise by adjusting the present supernatural quality and the data extraction module as delineated in [11].

III. PROPOSED WORK

The semantic indexing and concept based search can be achieved using DBpedia datasets. In the proposed work, a sematic application is developed using technologies such as DBpedia, Semantic indexing and OWL-DL. The entities related to search are retrieved from the DBpedia using DBpedia SPARQL endpoint. Then, the entities matching the user query are searched and retrieved by mapping the index with the ontology. The retrieved information is ranked in light of the orchestrating check. Finally, the results are appeared to the customer subject to the mentioning for rank decided for the pages.

a. DBpedia

DBpedia is a RDF outline which contains data ousted from Wikipedia and makes it available on the web. Using this, the requesting are killed Wikipedia and information can be separated. Data is gotten to using a requesting language called SPARQL. It revealed the majority of the information open in Wikipedia in a made structure. In like manner, furthermore the Dbpedia is a data base which tends to more than 2.9 million things including individuals, music collections, or motion pictures in 91 groundbreaking vernaculars.

Since, there are a few hindrances in catchphrase based sales search, thought based solicitation search is used which can be created by using DBpedia. Thusly, it is said to be semantic part search. DBpedia learning base has a few focal concentrations over the present data bases and it covers various spaces too.

b. SPARQL

SPARQL addresses SPARQL Protocol and RDF Query Language. It is a RDF question language. It recoups and controls educational record away in Resource Description Framework position (RDF). It takes after SQL, where SQL is used to recoup information from the database, SPARQL is used to recuperate information from the OWL records. It is used to request the cosmology and recuperate results from the OWL records subject to the customer's solicitation.

c. Overall Process

The following are the steps describing the overall flow of the system that has been adopted for soccer domain.

1. First step, the information is crawled from websites such as UFEA and SPORX since it is domain specific for soccer domain. This information is stored temporarily. It contains some basic information such as team players, goal, match venue and the match narrations etc in free-text format.
2. Using this basic information, the initial OWL files are populated.
3. From the initial OWL files, indexing is performed which consists of the basic information and the narrations.
4. From the initial OWL files, the content is enhanced and is used to populate the final ontology. Using this module, the extracted information such as fouls, off-sides, corners etc can be retrieved which is used to populate and get the final OWL files.
5. These OWL files are perused and listed to manufacture the last record.
6. In the next step, the data is retrieved from the ontology using the querying language SPARQL.
7. Once the data is retrieved using the queries, the indexing and the ontology are mapped using the similarity matching algorithm.
8. The retrieved results are ranked according to the ranking algorithm.
9. 9. Finally, the results are displayed to the user in the order of their rank calculated by the ranking algorithm. The overall architecture is shown in figure1.

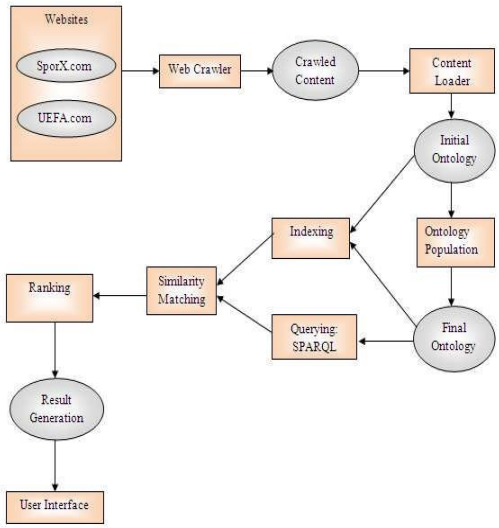

Figure 1. Overall Architecture

a. Website Crawling

Web crawler is fundamentally used to inspect the World Wide Web. This is cultivated for the most part to list the site pages. The crawler looks at the site pages subject to the meta-names present in the site pages source code. Those are the main data used to slither the site. The crawler visits the URLs and the majority of the hyperlinks are in like way visited, when the present page is finished breaking down. Additionally, the majority of the hyperlinks are visited until no new page is to be visited. The data ousted from the crawler is dealt with either as a database table or a substance record, and so forth.

b. Ontology plan

Cosmology expect a tremendous action in semantic web applications as it gives the information about the things truth be told. It besides draws in reusability and interoperability among various modules. The probability of supernatural quality structure, combination and re-trying give thought concerning hypothesis based data extraction and recovery [12].

For this framework, a focal cosmology is made, which has been utilized by each module of the structure particularly in data recovery and recovery structures. From the begin, magic has been made with the essential contemplations that are recovered by the web crawler. Besides, after that it is refined by improving its substance and fixing the issues to make a last predictable otherworldliness.

c. Content Enhancement

It is one of the most basic bits of hypothesis based semantic web applications. It is the course toward managing the unstructured data and changing over them into the dealt with data at last adding it to the information base. In this framework, the duty to the data extraction module would be the urgent data and normal language works assembled by slithering the site data[13]. The different leveled utilization of thinking for data recovery is inferred in [14].

d. Ontology masses

Otherworldliness masses is the way toward disengaging, orchestrating and managing learning by changing or mapping unstructured, semi-dealt with and formed information into cosmology people. In this development, an OWL file is made utilizing the data emptied during the past module. Introspective philosophy individuals isn't confined with the substances expelled from the IE module. As referenced beforehand, the crept information besides contains some key data. Thusly, the crept data is in like way used to make an OWL individual on the off chance that they don't exist in the learning base.

Cosmology individuals in addition wires including some real data about the match, for example, social events, experts, players, field, and so on. These data are populated to the otherworldliness by making an individual OWL record, which makes our last perspective.

e. Semantic Indexing

Mentioning can be performed from various perspectives, for example, semantic mentioning, catchphrase based mentioning, and so forth. Mentioning is an enormous bit of information recovery. It makes the mission for the information and the information recovery increasingly clear to discover and recover. Mentioning helps in speedier information recovery. In this structure, semantic mentioning is executed which is developed through lucene mentioning which gives unparalleled and is particularly expected to no closure substance mentioning. A comparative procedure is adjusted by [17-19].

f. Querying

There are different ways to deal with oversee demand a semantic knowledgebase and recover the outcomes, for example, watchword based, run of the mill language-based, see based and structure based semantic tending to. Among these strategies, watchword based recovery is intelligently satisfying for the end client to demand the data. Different systems, despite the manner in which that they engage continuously clear request to be organized, require more client facilitated exertion relying on the size of the area.

Notwithstanding the manner in which that catchphrase based interfaces have their very own extraordinary disturbs, for example, ambiguities, there are approaches to manage most extreme them. In this framework, SPARQL tending to language is utilized to demand the cosmology information base. This language gives tip top and looks like the SQL which is utilized to investigate the database. Q2 Semantic [16] used to locate the best sub framework conferring the request in the RDF chart.

a. Similarity Matching

Likeness sorting out or mapping would like to locate the semantic relationship between close segments of the perspective and the record. It is an inducing errand to accomplish semantic interoperability in structure the Semantic Web. To plot dissuading the once-over made, both the phonetic and the associate similitudes are pondered. In light of the mapping, the outcomes are recovered from the knowledgebase.

b. Ranking

Looking the web for negligible related data utilizing present solicitation structures acknowledges different harms like imprecision, monster thing, lacking to translate the conclusion of client's request, etc. To beat the impediments of the present approaches, a position based method has been proposed to update the centrality of once-over things. An arranging calculation is proposed to figure the circumstance of the records by utilizing the two elements: (I) the rehash of the watchword happened in the site page (ii) the commonplace factor of a similar catchphrase with the basic inquisitive words which are normally dismissed by the present pursue plans.

At long last, the got outcomes are appeared to the UI in the sales for their rank that has been managed by the arranging calculation. In this framework, catchphrase based tending to is utilized and the high recovery execution and adaptability is developed by semantic mentioning. To make reference to a couple, SPARK [15] utilizes a probabilistic solicitation arranging framework for structure up the best question tended to by the watchwords [5].

IV. RESULTS & DISCUSSIONS

A. Experimental Data

To evaluate this system, a few UFEA and SPORX matches and narrations have been crawled. Out of 1000 narrations, around 800 events have been extracted by this module. For the evaluation process, 15 queries are formed which are shown in Table 1.

TABLE 1. Queries for evaluation process

Query Number	Query
Q-1	Find all goals
Q-2	Find all goals scored by Manchester united
Q-3	Find all goals scored at Barcelona
Q-4	Find all goals scored by Messi at Barcelona
Q-5	Find all the Fouls
Q-6	Find all off-sides
Q-7	Find all punishments
Q-8	Find all the yellow cards received by Ronney
Q-9	Find all the defense players
Q-10	List all the forward position players
Q-11	Find all the saves done by the goalkeeper of Arsenal
Q-12	Find all the shoots delivered by the defense players of Liver pool
Q-13	List all the events involving Torres
Q-14	Find all the negative moves of Ronaldo
Q-15	Find all the matches won by Real Madrid

B. Parameters for Evaluation

The two critical parameters used to measure the display of request proposition are:

- Precision
- Recall.

Precision is resolved as the extent of relevant articles recuperated to the full scale things recouped. It is said to be the division of the reports recouped that are appropriate to the customer's information need and is resolved by the condition (4.1).

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \quad (4.1)$$

Survey is procured as the extent of material things recuperated to mean significant articles in the social occasion. It is in like manner said to be the piece of the reports that are imperative to the inquiry that are viably recouped and is resolved by the condition (4.2).

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \quad (4.2)$$

C. Result Analysis

The evaluation results show that the requesting and situating has made a solid improvement over its predecessor. In any case, when the request get progressively confounding, region unequivocal information extraction, standards and enlistment are relied upon to manage them. Along these lines, with this structure, a ton of chances are obliged the specialist to change their system as shown by the customer needs. Notwithstanding, the structure affirmations to give a comparative adaptability and usability. The precision regards which are resolved reliant on the amount of recouped and relevant files are showed up in table 2.

TABLE 2. Precision Calculation

Query	Retrieved documents	Relevant documents	precision
Q-2. Find all goals scored by Manchester united	8	4	0.5
Q-9. Find all the defense players	4	3	0.75
Q-10. List all the forward position players	6	5	0.83
Q-13 . List all the events involving Torres	5	3	0.6
Q-15. Find all the matches won by Real Madrid	8	6	0.75

The document inclusion between the recovered and the pertinent reports is demonstrated n Figure 2.

TABLE 3. Recall Calculation

Query	Total documents	Relevant documents	Recall
Q-2. Find all goals scored by Manchester united	10	4	0.4
Q-9. Find all the defense players	10	3	0.3
Q-10. List all the forward position players	10	5	0.5
Q-13 . List all the events involving Torres	10	3	0.3
Q-15. Find all the matches won by Real Madrid	10	6	0.6

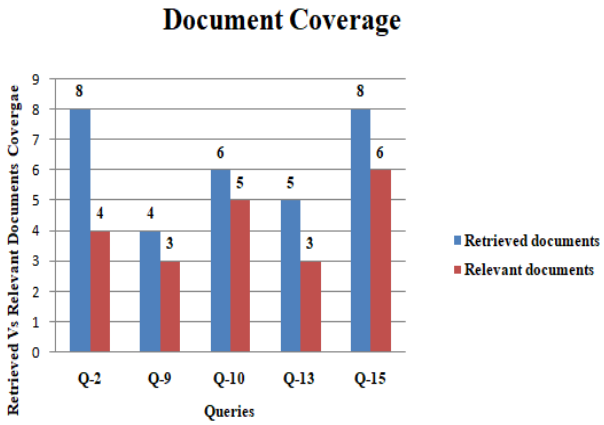

Figure 2. Document Coverage between Retrieved and Relevant Documents

The performance of this data recovery approach is estimated by utilizing precision and review proportions. The Precision proportion is determined as the proportion of important reports recovered to the all out records recovered relating to some example inquiries whose qualities are appeared in Figure 3.

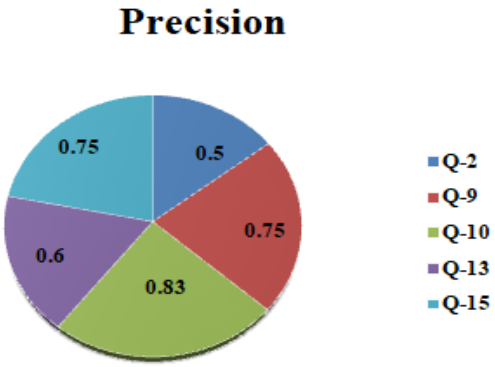

Figure 3. Precision Ratio for Retrieved and Relevant Document Coverage

The recall values which are calculated based on the complete number of archives and the important records which are recovered are appeared in table 3.

The document coverage between the total number of documents corresponding to the sample query and the relevant documents which are retrieved is shown in Figure 4.

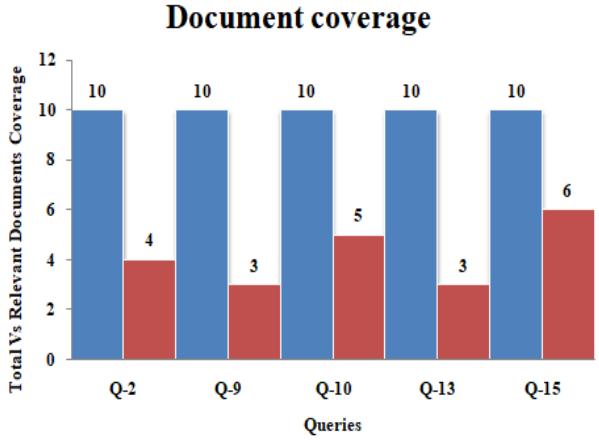

Figure 4. Document Coverage between Total and Relevant Documents

The recall ratio is determined as the proportion of significant reports recovered to the absolute records relating to some example questions whose qualities are appeared in Figure 5.

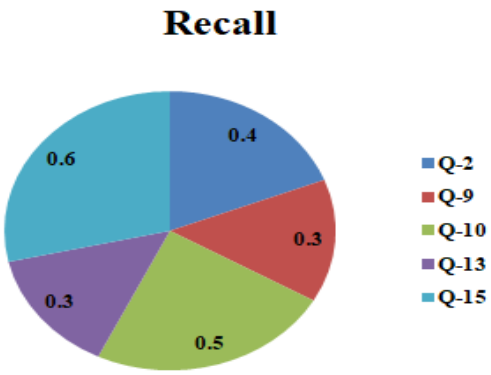

Figure 5. Recall Ratio for Total and Relevant Document Coverage

V. CONCLUSION

In this paper, a semantic based data recovery and an idea based way of thinking for data extraction for soccer space has been introduced. This paper combines the majority of the bits of semantic web explicitly Ontology, semantic standards, semantic mentioning, arranging and recovery. Precisely when this strategy is joined with catchphrase based intrigue interface, an easy to use, common and flexible semantic recovery framework is secured. In future, this way of thinking can be reached various spaces and besides novel arranging and mentioning figuring's can be related with improve the presentation of data recovery.

REFERENCES

1. Soner Kara, Ozgur Alan, Orkunt Sabuncu, Samet Akpinar, Nihan K. Cicekli, Ferda N. Alpaslan (2012), 'An ontology-based retrieval system using semantic indexing', Information Systems, vol. 37, pp. 294–305.
2. Sajendra Kumar, Ram Kumar Rana, Pawan Singh (2012), 'Ontology based Semantic Indexing Approach for Information Retrieval System', International Journal of Computer Applications, vol. 49, pp. 0975 – 8887.
3. Licia Sbattella, Roberto Tedesco (2013), 'A novel semantic information retrieval system based on a three-level domain Model', The Journal of Systems and Software, vol. 86, pp. 1426– 1452.
4. Mauro Dragoni , Célia da Costa Pereira, Andrea G.B. Tettamanzi (2012), 'A conceptual representation of documents and queries for information retrieval systems by using light ontologys', Expert Systems with Applications, vol. 39, pp. 10376–10388.
5. Sheetal A. Takale, Sushma S. Nandgaonkar (2010), 'Measuring Semantic Similarity between Words Using Web Documents', (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1.
6. P.Priya, Dr.R.R.Rajalaxmi (2013), 'Semantic Entity Suggestion for Effective Search using DBpedia', Proceedings of the International Conference on Intelligent Instrumentation, Optimization and Signal Processing.

7. V.Uren, Y.Lei, V.Lopez, H.Liu, E.Motta, M.Giordanino (2007), 'The usability of semantic search tools: a review', Knowledge Engineering Review, vol. 22, pp. 361-377.
8. A. Schutz, P. Buitelaar (2005), 'Relext: a tool for relation extraction from text in ontology extension', in: The Semantic Web (ISWC), pp. 593–606.
9. N. Kiyavitskaya, N. Zeni, J.R. Cordy, L. Mich, J. Mylopoulos (2009), 'Cerno: light-weight tool support for semantic annotation of textual documents', Data and Knowledge Engineering, vol. 68, pp. 1470–1492.
10. A. Wessman, S. W. Liddle, D. W. Embley (2005), 'A generalized framework for an ontology-based data-extraction system', proceedings of the 4th International Conference on Information Systems Technology and its Applications, pp. 239– 253.
11. D. Tunaoglu, O. Alan, O. Sabuncu, S. Akpinar, N.K. Cicekli, F.N. Alpaslan (2009), 'Event extraction from Turkish football web-casting texts using hand-crafted templates', Proceedings of the IEEE International Conference on Semantic Computing, ICSC '09, pp. 466–472.
12. D. Oberle, A. Ankolekar, P. Hitzler, P. Cimiano, M. Sintek, M. Kiesel, B. Mougouie, S. Baumann, S. Vembu, M. Romanelli, P. Buitelaar, R. Engel, D. Sonntag, N. Reithinger, B. Loos, H.-P. Zorn, V. Micelli, R. Porzel, C. Schmidt, M. Weiten, F. Burkhardt, J. Zhou, DOLCE ergo (2007), 'SUMO: on foundational and domain models in the SmartWeb integrated ontology (SWIntO)', Journal of Web Semantics, vol. 5, pp. 156–174.
13. M. Liao, A. Abecker, A. Bernardi, K. Hinkelmann, M. Sintek (1999), 'Ontologies for knowledge retrieval in organizational memories', Proceedings of the Workshop on Learning Software Organizations, Fraunhofer Institute for Experimental Software Engineering, pp. 11–25.
14. H.-M. Muller, E.E. Kenny, P.W. Sternberg (2004), 'Textpresso: an ontology based information retrieval and extraction system for biological literature', PLoS Biology, vol. 2, pp. 309.
15. Q. Zhou, C. Wang, M. Xiong, H. Wang, Y. Yu(2007), 'Spark: adapting keyword query to semantic search', Proceedings of the 6th International Semantic Web Conference and Second Asian Semantic Web Conference (ISWC/ASWC2007), Busan, South Korea, vol. 4825, pp. 687–700.
16. H. Wang, K. Zhang, Q. Liu, T. Tran, Y. Yu (2006), 'Q2 semantic: a lightweight keyword interface to semantic search', in: ESWC, Springer, pp. 584–598.
17. U. Shah, T. Finin, A. Joshi, R.S. Cost, J. Matfield (2002), 'Information retrieval on the semantic web', Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM, New York, NY, USA, pp. 461–468.
18. J. Davies, R. Weeks (2004), 'Quizrdf: search technology for the semantic web', Proceedings of the Hawai International Conference on System Sciences, vol. 4, pp. 40112–40120.

AUTHORS PROFILE

P.Priya Assistant Professor, Department of Computer Science and Engineering BIET, Hyderabad, Telangana – 501510
Email id: priyaponnusamy29@gmail.com