iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4295

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Hyponym Relation Extraction from Hyperlinks using Motif Patterns with Feature Combination

Ms.P.Muthumani[1], Mr. R. Karthik[2]
[1]*PG Scholar,* [2]*Assistant Professor, Department of Computer Science and Engineering, Vivekanandha College of Technology for Women, Elayampalayam, Thiruchengode. Tamilnadu, India.*

Abstract: ***The project presents the grammatical relationship between concepts in Knowledge Graphs such as Word-Net and DBpedia. Past work on grammatical relationship method has either the form grammatical system connection between concepts. To develop the grammatical relationship method, namely w path, to connect those two ways, using Information Content to weight the abridged path distance between ideas. Every collection-based Information Content is computed from the allocation ideas over a textual collection. Weighing the grammatical relationship between the communication is an important action in modify processing field. The project planned a new relationship measure. So, the new hyponym relation extraction approach based on the network pattern of Wikipedia hyper-sonic was planned. Every hyper-sonic was mapped 13 types of three-node patterns.***

Keyword: Semantic similarity, WordNet, DBpedia, Information Content, Knowledge Graphs, Hyponym relation, Three-node motifs, 13-dimensional vector

I. INTRODUCTION

Big data is a terminology for data sets that are so larger or difficult that traditional data processing application software is inadequate to deal with them. Today, many organizations are collecting, storing, and analyzing massive amounts of data. This data is regularly referred to as “Big Data” because of its volume, the velocity with which it arrives and is used to enhance decision making. Provide understanding and verification, support and modify processes.

Big data challenges include data storage, data analysis, search, dividing, transfer, visualization, querying, rejuvenate and information privacy. Big data basically includes following three steps:

1) *Volume* – collect data from a variety of sources
2) *Velocity* – speed (data generated fast and need to be processed fast)
3) *Variety* - it completes missing pieces through the data fusion.

II. RELATED WORKS

1) *Saeedeh Shekarpour et al [2]* describes an extension of a semantic interpretation of user queries for question answering on interlinked data which contains both duplicated and fragmented information on a large number of domains. One way to enable non-experts users to access this data compendium is to provide keyword search frameworks that can capitalize on the inherent characteristics of Linked Data. Developing such systems is challenging for three main reasons. First, resources across different datasets or even within the same dataset can be homonyms. Second, different datasets employ heterogeneous schemas and each one may only contain a part of the answer for a certain user query. In conclusion, developing a join together formal query from keywords across the variety of dataset requires abuse links between the different dataset on both schema and context levels.
2) *Ioana Hulpus set al [3]* describe a path based well-formed relatedness on linked data and its use to word and entity disambiguation. There are many approaches dealing with both problems but most of them rely on word or concept distribution over Wikipedia. They are therefore not applicable to concepts that do not have a rich textual description. In this paper, we tackle two strongly interdependent problems, semantic relatedness, and disambiguation. The aim of semantic relatedness is to weight the semantic gang between pairs of concepts. The aim of entity and word sense disambiguation is to link strings in the text to the related concepts in superficial knowledge bases. The rationale behind the previous relatedness measure is that the more and shorter relation paths between two nodes, the higher their relatedness. However, it has been long known that not all direct relationships weight the same. Manual assignment of weights based on relationship type is infeasible, given a lot of relationship types in knowledge graphs. In this paper, developed a unique measure of the strength of relations in knowledge graphs, called relation rarity. Use this measure for computing semantic relatedness as well as same.

3) *Xiang Lian et al [4]* describes the same join processing on the uncertain data streaming environment has many practical applications such as sensor networks, object observing and monitoring, and so on. Previous works usually assume that stream processing is conducted over precise data. In this paper, important problems of parity join processing on stream data that congenitally contain uncertainty, where the incoming data at each time stamp is uncertain and imprecise. To tackle the challenge with respect to efficiency and effectiveness such as limited memory and small response time. Join processing over uncertain data streams is useful in many practical applications such as data cleaning outlier detection and object tracking and monitoring. In the application of child care or elder care, the positions of both children/elders and their accompanied persons can be captured or identified by sensing devices such RFID readers or GPS. In order to avoid the situation where children or elders wander too far away, an alert about this event to the nanny/helper would be helpful to take good care of the children and elders. That compares (i.e., join) real-time trajectories of both persons that are in streaming fashion. In this case, where two persons have positions far away from each other for some period of time, the number of matching pairs between the two data streams would greatly decrease, which will trigger the alert of an abnormal event and report it to nanny/helper immediately.
4) *Ye Yuan et al [5]* describes an efficient keyword search on uncertain graph data. As a popular search has been applied to retrieve useful data in documents, texts, graphs, and even relational datasets. Anyhow, there is no task on watchword search over fitful graph data even though the uncertain graphs have been widely used in many real applications on networks. Following the similar answer definition for keyword search over deterministic graphs, consider a subtree in the uncertain graph as an answer to a keyword query if 1) it contains all the keywords; 2) it has a high score based on keyword matching; 3) it has low uncertainty. Keyword search over deterministic graphs is already a hard problem. In this paper, focus on threshold-based probabilistic keyword search (T-KS) over large uncertain graph data. An uncertain graph g with each node attached some text, i.e., node eight containing two keywords {b, c}. A real number associated with each edge represents the existence probability of edge. A Possible World Graph (PWG) of an uncertain graph is a possible instance of the uncertain graph. It contains all nodes and a subset of edges of the uncertain graph, and its existence probability is the product of the probabilities of all the edges it has. To avoid the hard problem, greedy algorithms are used to find the approximate top-k results.
5) *Michael Schuhmacher et al [6]* describes a knowledge-based graph document modeling, a graph-based semantic model for representing document content. This method relies on the use of a semantic network, namely the DBpedia knowledge base, for acquiring fine-grained information about entities and their semantic relations, thus resulting in a knowledge-rich document model. To demonstrate the benefits of these semantic representations in two tasks; entity ranking and computing document semantic similarity. Entity ranking is the task of ordering a given set of entities on the basis of their relevance with respect to a specific reference entity. This ranking task has the advantage that it provides a focused, extrinsic evaluation of our different weighting methods: besides, there exists established gold standard datasets against which compare our approach. Entity ranking can be seen as similar in spirit to computing word relatedness. Given an input text document, first, sanctify it by identifying the set of concepts it contains. To this end, words and phrases are illustrated with DBpedia concepts using a document entity linking system, e.g., DBpedia Spotlight. Given a mention and its candidate entities, the entity linker finds its most likely meaning in context.

III. METHODOLOGY

In this project, previous techniques on uncertain query processing or the skyline computation cannot directly tackleour complex pg-KWS problem, which involves both keyword search queries and probabilistic RDF graphs. Therefore, one straightforward method is illustrated as follows. For each probabilistic r-radius graph extracted from probabilistic RDF graph, we first calculate its 2D feature/entropy scores online query keywords. Then, retrieve those probabilistic r-radius graphs that are not dominated by others in the 2D score space.

Since query keywords are online specified, this method has to compute scores of probabilistic r-radius graphs for every query from scratch and perform the dominance check for all graphs, which involves an exponential number of possible worlds and it is very inefficient. In contrast, our pg-KWS approach applies effective pruning techniques to reduce the search space, and the number of subgraph candidates is much smaller than that of the straightforward method. Thus, pg-KWS approach, below, we will not plot curves for the straightforward method. In addition evaluate the performance of our pg-KWS approach, in terms of the CPU time and I/O cost, where the CPU time is the time cost of accessing the index and applying pruning methods, and the I/O cost is the number of page accesses during the index traversal.

A. Document Selection

A list of Wikipedia articles is downloading and selected i.e., copied to our folder using this module.

B. Wag Construction

This module constructs the Wikipedia Article Graph. During this, the pages referred to external web pages or files not found in this collection are eliminated. Using regular expression, the pages are found out using the links. Then a graph is formed such that the pages are being nodes and the links are being edges. This considers the topological structure of Wikipedia hyperlinks as an important type of feature in hyponym relation extraction. Each Wikipedia article page represents a domain-specific term. It contains a number of hyperlinks pointing to other article pages. Fig. 1 shows a fragment of an article page k-medoids algorithm, which contains three hyperlinks. These hyperlinks and article pages can be considered as directed graphs. Then these graphs are named as the Wikipedia article graphs (WAGs). The hyperlinks in a WAG can imply semantic relations such as hyponym relation between the two connected article pages.

The k-medoids algorithm is an aggregate algorithm related to the k-means algorithm and the medoids-shift algorithm. Both the k-medoids algorithms are partitioned and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses data points as centers

Fig.1 Fragment of a Wikipedia article page that illustrates the semantics of Wikipedia hyperlinks.

For example, in “Fig. 1” the first hyperlink indicates that K-medoids algorithm is a hyponym of a clustering algorithm, and the second hyperlink indicates that K-medoids algorithm is a co-hyponym of k-means.

C. Data Cleaning Module

1) *Add Stem Word:* In this module, the word and its stem word is keyed in and saved into the table. The details are saved in ‘Stemword’ table.
2) *Add Stopword:* In this module, the stop word is keyed in and saved into the table. The detail is saved in ‘Stopword’ table.
3) *Add Synonym Word:* In this module, the word and its synonym word is keyed in and saved into the table. The details are saved in ‘Synonym’ table.
4) *Add Hyponym Word:* In this module, the word and its hyponym word is keyed in and saved into the table. The details are saved in ‘hyponym’ table.

D. Preprocessing

In this module, all the documents downloaded are applied with stemming, stop word removal and synonym word replacement.

E. Motif Pattern Construction

In this module, the three node network motif patterns are constructed. The two parameters below were utilized to qualify the three-node motifs.

1) Z-Score indicates the statistical significance of a network motif. The Z-Score of motif j is formally defined in (1).

$$\text{Z-Score}(j) = \frac{N(j) - \overline{Nr(j)}}{\sigma(Nr(j))} \quad (1)$$

Where N(j) is the number of occurrences of motif j($1 \leq j \leq 13$) in network N.Nr(j)is the average number of occurrences of motif j in an ensemble of randomized networks with the same degree of distribution as network N.r(j) is the standard deviation of Nr(j). in general, a motif with a high Z-Score indicates that the motif appears in a particular network (N) more frequently than in randomized networks.

2) A new parameter, Hyponym Hyperlink Rate (HHR), was introduced to describe the sparsity of hyponym relations within a network motif. The HHR of motif j is defined in (2). The higher the HHR of a network motif is, the denser the hyponym hyperlinks in the motif are. The condition means that if a hyperlink appears in a motif with high HHR, then this hyperlink is likely to be a hyponym hyperlink.

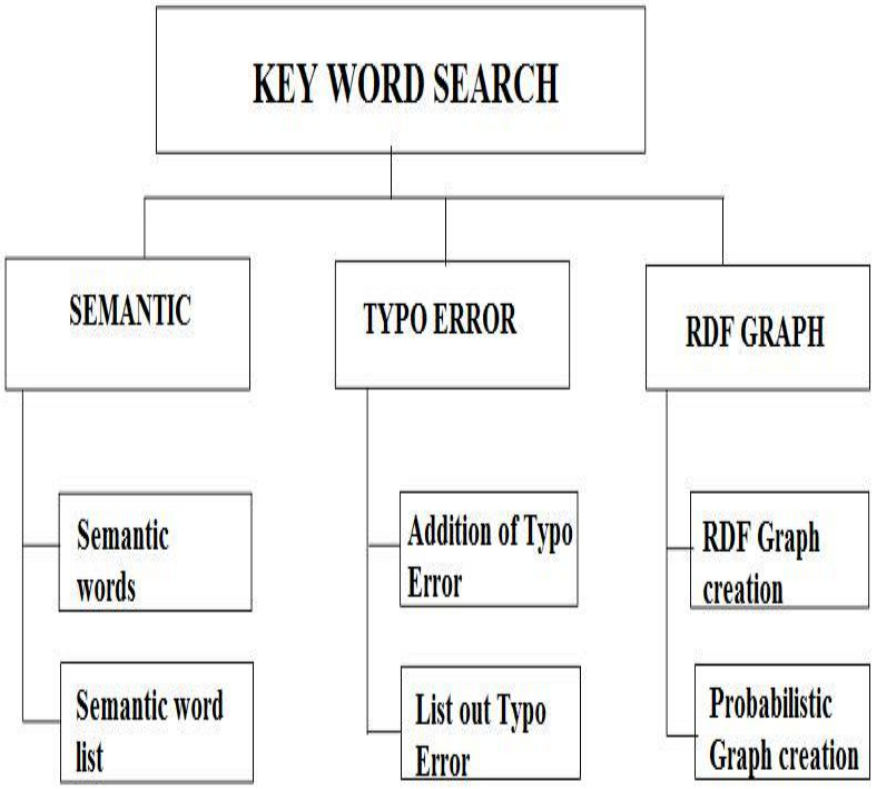
$$HHR(j) \quad (2)$$

F. Feature Based Text Content

In this module, the features based on text content are also combined. For example, the word mobile if contained, it relates the links of pages containing the <company name> mobile related phrases even f does not behave as hyponym for the same context

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

IV. SYSTEM DIAGRAM



V. CONCLUSION

In present work, the text-search conditions are limited to keywords. However, IR systems usually offer a richer repertoire of predicates: phrase matching (i.e., several contiguous keywords), proximity search (i.e., several non-contiguous keywords within short distance), negated conditions (i.e., taboo words that must not appear in a result), query expansion (i.e., adding related words that are not explicitly given in the query), and more.

A novel keyword search based framework is geared toward this and can compute meaningful rankings even for such richer queries with a structured "backbone". However, efficient indexing and query processing pose major challenges.

Typo Errors are also considered so that the document data set if contains misspelled words they are also replaced with correct values. The number of ‘Subjects ‘ are related to a given ‘Object' word during the RDF graph construction. In addition, the probability of each ‘Subject’ forgiven ‘Object’ is also tracked. Likewise, a probability threshold is applied so that values above the threshold are filtered and subgraph is shown for the given query words. Data can be processed very easily.

1) Data preprocessing steps as Stemming, stop words removal and synonym word replacement is also considered.
2) This approach may work well in a domain where the hyponym relations among domain-specific terms are containing more different words for same meaning.
3) The features based on text content are also combined. For example, the word mobile if contained, it relates the links of the page contained the < company name > mobile and other mobile related phrases even if does not behave as hyponym for the same context.

VI. RESULTS

A. Hit Rate Analysis

The following “Table 6.1” describes experimental result for a number of query search process in existing and proposed hit rate analysis. The table contains a number of the search query, existing hit rate and proposed hit rate details are shown.

S. No.	Number of Query Search	Existing System Hit Rate	Proposing System Hit Rate
1	25	0.265	0.313
2	50	0.278	0.285
3	75	0.312	0.321
4	100	0.345	0.349
5	125	0.387	0.392
6	250	0.394	0.98
7	275	0.404	0.408
8	300	0.431	0.437

Table 6.1 Performances Analysis- Hit Rate

The following "Figure 6.1" describes experimental result for number of query search process in existing and proposed hit rate analysis. The table contains a number of the search query, existing hit rate and proposed hit rate details are shown.

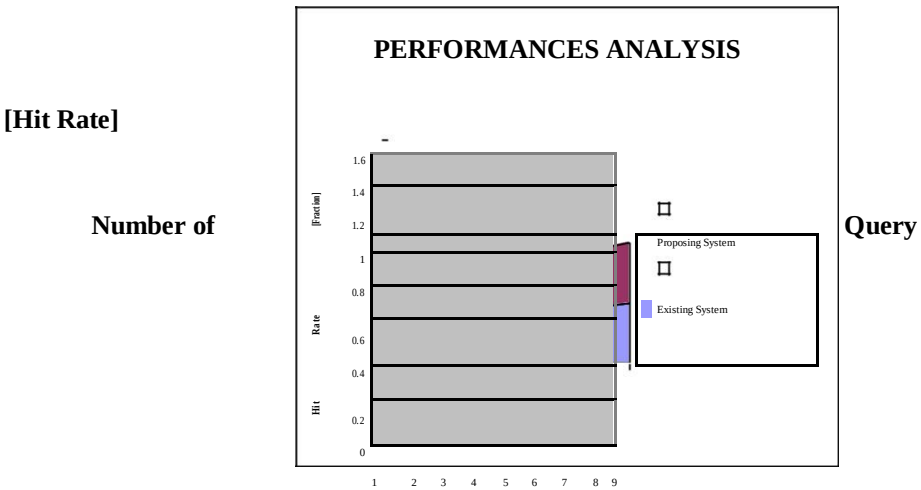


Fig 6.1 Performances Analysis- Hit Rate

B. Average Delay Analysis

The following "Table 6.2" describes experimental result for a number of query search process in existing and proposed average delay of query analysis. The table contains a number of the search query, existing hit average delay, and proposed average delay details are shown.

S. No.	Number of Query Search	Existing System Average Delay	Proposing System Average Delay
1	25	60.22	62.12
2	50	63.54	65.04
3	75	70.13	75.31
4	100	74.34	78.46
5	125	79.66	84.37
6	250	83.75	86.79
7	275	87.39	89.87
8	300	91.67	92.08

Table 6.2 Performances Analysis- Average Delay

The following "Figure 6.2" describes experimental result for number of query search process in existing , and proposed average delay of query analysis. The table contains a number of search query, existing hit average delay and proposed average delay details are

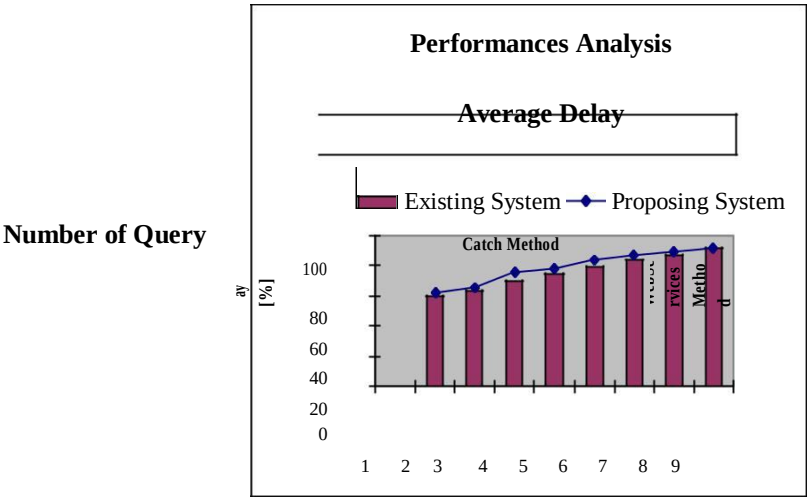


Fig 6.2 Performances Analysis- Average Delay

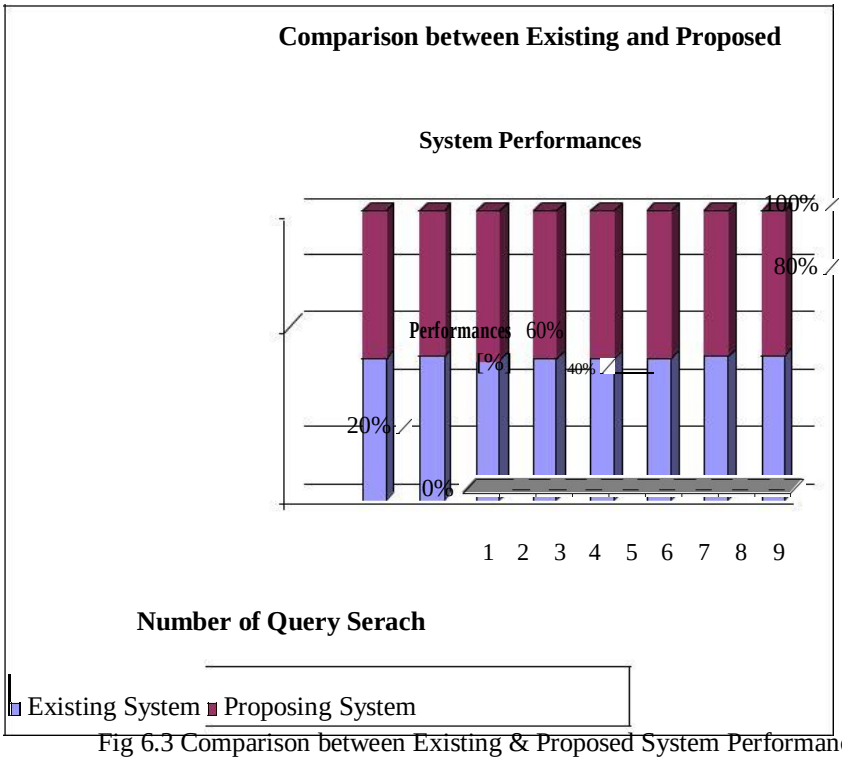


Fig 6.3 Comparison between Existing & Proposed System Performances

The above "Figure 6.3" describes experimental result for a number of query search process in related technology in the system and hit rate analysis. The table contains a number of search query, related method and, existing hit rate and proposed hit rate details are shown.

VII. FUTURE ENHANCEMENTS

This project shows the growing importance of introducing new keyword search mechanisms for RDF data. The user experience in querying an RDF graph can be significantly improved by exploiting different criteria for ranking query answers towards meeting the user information needs. The problem of exploring how to combine various subjects with a given object is solved now. Research progress in this area does not necessarily demand to work from scratch; a different point of view on how to employ or adapt existing algorithms and techniques are considered. In future extensive experiments may be conducted to verify the effectiveness and efficiency of our proposed approaches. The following enhancements can be made in future.

1) The application if developed as web services, then many applications can make use of the records.

2) In future work, want to combine the keyword search with scalable approaches for construct RDF graph and to deliver results even faster.

REFERENCES

[1] Ganggao Zhu and Carlos A. Iglesias," Computing Semantic Similarity of Concepts in Knowledge Graphs" vol. 29, no. 1, january 2017.

[2] Shekarpour. S, E. Marx, A.-C. N. Ngomo, and S. Auer, "Sina: Semantic interpretation of user queries for question answering on interlinked data," Web Semantics: Sci. Services AgentsWorld Wide Web, vol. 30, pp. 39–51, 2015.

[3] Hulpus.I, N. Prangnawarat , and C. Hayes, "Path-based semantic relatedness on linked data and its use to word and entity disambiguation," in Proc. 14th Int. Semantic Web Conf., 2015, pp. 442–457.

[4] Xiang Lian, and Ye Yuan "Similarity Join Processing On Uncertain Data Streams" in Proc. 7th Int. Semantic Web Conf., 2015, pp. 442–457, 2010.

[5] Ye Yuan, A. Raganato, and R. Navigli, "Efficient Keyword Search On Uncertain Graph Data" Trans. Assoc. Comput. Linguistics, vol. 2, pp.231–244, 2013.

[6] Schuhmacher.M and S. P. Ponzetto, "Knowledge-based graph document modeling," in Proc. 7th ACM Int. Conf. Web Search Data Mining, 2014, pp. 543–552.

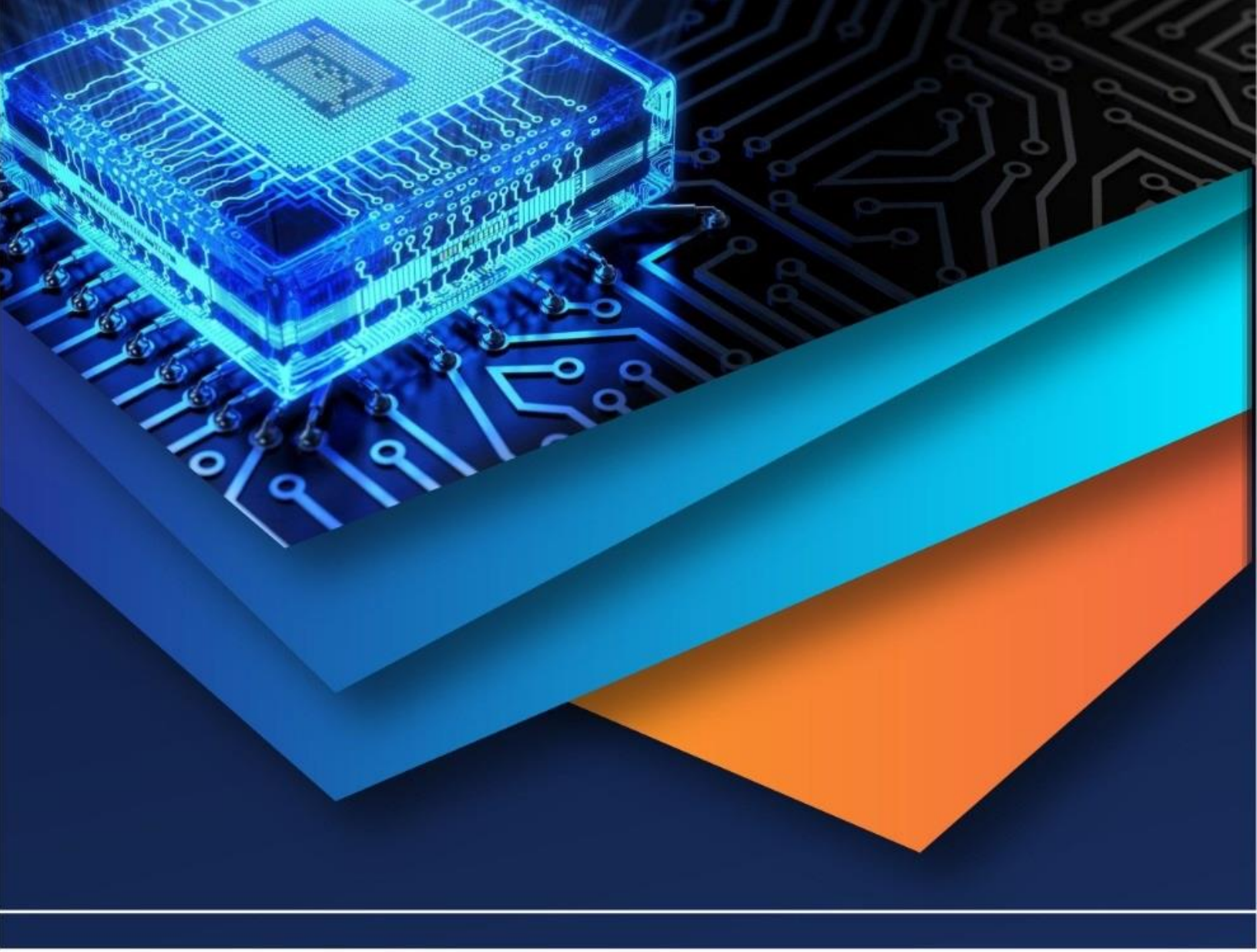
[7] Bizer.C, "DBpedia-a crystallization point for the web of data," Web Semantics: Sci. Services Agents World Wide Web, vol. 7, no. 3, pp. 154–165, 2009.

[8] Dong. X. L, L. Berti-Equille, and D. Srivastava, "Integrating conflicting data: The role of source dependence,"Proc. VLDB Endowment, vol. 2, no. 1, pp. 550–561, 2009.

[9] Hoffart.J, F. M. Suchanek, K. Berberich, E. Lewis-Kelham,G. de Melo, and G. Weikum, "Yago2: exploring and querying world knowledge in time, space, context, and many languages," in Proc. 20th Int. Conf. Companion World Wide Web, 2011, pp. 229–232

[10] Navigli.R ,E. Hovy, and S. P. Ponzetto, "Collaboratively built semi-structured content and artificial intelligence: The story so far," Artif. Intell., vol. 194, pp. 2–27,2013.

[11] Hovy.E, and S. P. Ponzetto, "Babelnet: The automatic construction, evaluation, and application of a wide-coverage multilingual semantic network," . Intell., vol.193,pp.217–250,20.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY