iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 Issue: XII Month of publication: December 2017

DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

Extracting the Attributes of Entities from Semi Structured Information Using XSearch

S.Nagarajan[1], Dr.K.Perumal[2]
[1]*Department of Computer Applications, School of Information Technology, Madurai Kamaraj University, Madurai.*
[2]*Associate Professor, Department of Computer Applications, School of Information Technology,*

***Abstract:** The rapid progress of network and storage technologies has led to a huge amount of electronic data such as web pages and XML data has been available on intra and internet. More than 80% of today's data is composed of unstructured or semi structured data. Various techniques are available to extract useful data from the webpage or XML documents. These electronic data are heterogeneous collection of ill-structured data that have no rigid structure, and are often called semi-structure data. These semi-structured data are stored in large repositories (XML databases) and stored as a graph internally in database with tuple as nodes and relationships as edges. As there is ever-growing availability of semi-structured information on web and digital libraries, there is a need of effective keyword search in order to fetch the correct and proximal result on Semi-Structured Data. The data can be extracted from the semi structured documents such as web pages and XML documents is still difficult because it has no the proper structure. This paper focuses on the semi structured data that can be extracted from the webpage or XML documents using the keywords search. The keyword search method is the effective method to search the documents in web pages or XML documents. In this paper, the XSearch method is proposed to search a keyword. After searching, the list of relevant semi structured text is viewed. In this paper, the Artificial Neural Network(ANN) is used to extract the information from XML documents and can be modeled as graph. In Neural Network the nodes act as entities. The entities, attributes and their relationships are extracted from the semi structured data using ANN. Compare to the existing keyword search methods such as SLCA, MLCA, the XSearch method is the most effective and it achieves highest accuracy and F- measure.*

***Keywords:** Keywords: Semi structured, XSearch, Data Mining, Information Extraction, Information retrieval, ANN*

I. INTRODUCTION

The Web provides access to a large number of information sources of all kinds. The major models for semi-structured data exchange over distributed information sources are the Object Exchange Model, the Extensible Markup Language and the Resource Description Framework (RDF). Structured query language like XPath and XQuery is used to search XML data in XML repository, the relevant keyword search on semi structured data is challenging task because each result fetched can have multiple matches. The speeding up of query processing is achieved using languages such as XQuery or XPath, to achieve improvements on efficiency in xml keyword search is done by using indexes and materialized views. The query processing speed can be increased by using cache, it stores the results of previously answered queries in order to answer succeeding queries faster by reusing these results. The different approaches for using caches are, approach checks whether or not a current query Q can be directly answered from the result of a previously answered query Qi stored in the cache. The new query is otherwise submitted to the source (xml database).

Data Mining is concerned with the discovery of patterns and relations in large collection of data. Data Mining is referred to as Knowledge Discovery in Databases. It deals with issues such as representation schemes for the concept or pattern to be discovered, design of appropriate functions and algorithms to find patterns. Most of the data mining algorithms can handle data with a fixed structure, where data scheme is defined in advance. However, data on the web and bioinformatics databases often lack such a regular structure. We call such data as semi-structured. By rapid progress of network and storage technologies, a huge amount of electronic data such as Web pages and XML data has been available on intra and internet. These electronic data are heterogeneous collection of ill-structured data that have no rigid structure, and are often called semi-structured data

Structured data is one that can be neatly modelled, organized, and formatted into ways that are easy for us to manipulate and manage. The most frequent examples include databases, spreadsheets, fixed-format files, log files, etc. Unstructured data incorporates the mass of information that does not fit easily into a set of database tables. The most recognizable form of unstructured data is text in documents, such as articles, slide presentation or message components of e-mails.

Semi-structured data refers to set of data in which there is some implicit structure that is generally followed, but not enough of a regular structure to qualify for the kinds of management and automation usually applied to structured data. Examples include the

World Wide Web, bioinformatics databases and data ware housing. Unlike unstructured raw data such as image and sound, semi-structured data has some structure: objects share (parts of) their structure.

Despite the structural irregularity, semi-structured data typically do possess some structure [3]. Such structures implicit in semi-structured data can serve the following purposes: optimizing query evaluation, obtaining general information contents, facilitating the integration of data from several information sources, improving storage, assisting in building indexes and views and making it possible for structure-based document clustering [4][5]. Most data mining algorithms are not designed for semi-structured data and should at least be adapted in order to deal with such data [1].

II. SEMI-STRUCTURED DATA

The growth of the use of semi-structured data has created new opportunities for data mining, which has traditionally been concerned with tabular data sets, reflecting the strong association between data mining and relational databases. XML, being the most frequent way of representing semi-structured data, is able to represent both tabular data and arbitrary trees. Any particular representation of data to be exchanged between two applications in XML is normally described by a schema often written in XSD. Practical examples of such schemata, for instance News ML, are normally very sophisticated, containing multiple optional sub trees, used for representing special case data. Frequently around 90% of a schema is concerned with the definition of these optional data items and sub-trees. Messages and data, therefore, that are transmitted or encoded using XML and that conform to the same schema are liable to contain very different data depending on what is being transmitted. XPath is the standard mechanism used to refer to nodes and data items within XML. It has similarities to standard techniques for navigating directory hierarchies used in operating systems user interfaces. To data and structure mine XML data of any form, at least two extensions are required to conventional data mining. Such data presents large problems for conventional data mining.

The use of semi-structured data can be felt in the areas involving raw data which does not have any fixed format. Semi-structured data is convenient for data integration. Web-sites containing semi-structured data are ultimately graphs. More and more data sets do not fit in the rigid relational model because the individual data items do not have the same structure completely. Rather, the data items share only partly the same structure. Such databases are called semi-structured databases [5]. In a semi-structured database, there is no fixed database schema: conceptually the data is stored in a graph-like structure (like XML) which contains both information about the data as well as the data itself. Prime examples of semi-structured databases are XML databases and many of the bioinformatics databases.

A. *Structured Model*

The semi-structured model is a database model where there is no separation between the data and the schema, and the amount of structure used depends on the purpose. The advantages of this model are the following:

1) It can represent the information of some data sources that cannot be constrained by schema.
2) It provides a flexible format for data exchange between different types of databases.
3) It can be helpful to view structured data as semi-structured (for browsing purposes).
4) The schema can easily be changed.
5) The data transfer format may be portable.

III. BACKGROUND

This section provides an overview on XML data models, query models and the definitions of query results. XML,[2] other markup languages, email, and EDI are all forms of semi-structured data.OEM (Object Exchange Model)[3] was created prior to XML as a means of self-describing a data structure. XML has been popularized by web services that are developed utilizing SOAP principles.

Some types of data described here as "semi-structured", especially XML, suffer from the impression that they are incapable of structural rigor at the same functional level as Relational Tables and Rows. Indeed, the view of XML as inherently semi-structured (previously, it was referred to as "unstructured") has handicapped its use for a widening range of data-centric applications. Even documents, normally thought of as the epitome of semi-structure, can be designed with virtually the same rigor as database schema, enforced by the XML schema and processed by both commercial and custom software programs without reducing their usability by human readers.

In view of this fact, XML might be referred to as having "flexible structure" capable of human-centric flow and hierarchy as well as highly rigorous element structure and data typing. The concept of XML as "human-readable", however, can only be taken so far. Some implementations/dialects of XML, such as the XML representation of the contents of a Microsoft Word document, as implemented in Office 2007 and later versions, utilize dozens or even hundreds of different kinds of tags that reflect a particular problem domain - in Word's case, formatting at the character and paragraph and document level, definitions of styles, inclusion of citations, etc. - which are nested within each other in complex ways. Understanding even a portion of such an XML document by reading it, let alone catching errors in its structure, is impossible without a very deep prior understanding of the specific XML implementation, along with assistance by software that understands the XML schema that has been employed. Such text is not "human-understandable" any more than a book written in Swahili (which uses the Latin alphabet) would be to an American or Western European who does not know a word of that language: the tags are symbols that are meaningless to a person unfamiliar with the domain.The XML documents represent hierarchically structured information and are generally modeled as Ordered Labeled Trees. XML document can be modeled as a tree, which is labeled and directed. Each element, attribute and text value in the XML document is a node in the XML tree. Nodes represent XML elements and are labeled with corresponding element tag names, organized following their order of appearance in the document. Each edge in the XML tree represents the membership of the element corresponding to the child node, under the element corresponding to the parent node in the XML document. Graph model models an XML document as a graph. The query models, XSearch[8] is structured query format, each query term has the format of l: k where l and k are keywords. There is another way to help non-expert users for querying XML documents is graphical query environment. Visual query processing DVQ[6] for XML database systems is used to displays the structure of the XML data to the user and XQBE [1] graphical environment to query XML data on web repository. A query result [4] of a keyword search on an XML document or search on XML repository retrieves a sub tree or subgraph as a result for keyword search, which contains both relevant keyword matches i.e., the XML nodes that match the keywords and other nodes that are relevant by the search engine and XSeek is the first system that automatically infers relevant non-matches i.e. XSeek outputs the subtree rooted at each return node as relevant non-matches. The information retrieval models [12] are divided into set theoretic models, algebraic models and probabilistic models. The Boolean model, case-based reasoning model, fuzzy set model and extended Boolean model are four main types in Set theoretic models. An algebraic model contains vector space model, generalized vector space model, latent semantic indexing model and neural network model. A probabilistic model includes probabilistic model, inference network model and brief network model. The issues are lack of semantics mainly in the Boolean model and the issues in algebraic models are maintaining difficulty, computational cost and lack of validation. The primary issues in probabilistic models are difficult to implement and computing cost.

IV. INTRODUCTION OF NEURAL NETWORKS

An Artificial Neuron is basically an engineering approach of biological neuron. It has device with many inputs and one output. ANN is consist of large number of simple processing elements that are interconnected with each other and layered also

A. Characteristics Of Neural Networks

The Characteristics are basically those which should be present in intelligent System like robots and other Artificial Intelligence Based Applications[5].An NN can formally be defined as: a massively parallel interconnected network of simple (usually adaptive) processing elements which is intended to interact with the objects of the real world in the same way as biological systems do. NNs are designated by the network topology, connection strength between pairs of neurons (called weights), node characteristics, and the status updating rules. Normally, an objective function is defined which represents the complete status of the network and the set of minima of it corresponds to the set of stable states of the network. NN-based systems are usually reputed to enjoy the following major characteristics: generalization capability, adaptivity to new data/information, speed due to massively parallel architecture, robustness to missing, confusing, ill-defined/noisy data, and capability for modeling nonlinear decision boundaries. NNs have been applied, so far, to the tasks like IR, IE, and clustering (self organization) of web mining[3], and for personalization.

B. Information Retrieval (IR)

IR is the automatic search of the relevant information contained in a set of knowledge, guaranteeing at the same time that non-relevant retrieved information is as less as possible. The aim must be to reach an improvement in retrieval results according to two key concepts in IR: recall and precision. Recall bears in mind the fact that the most relevant objects for the user must be retrieved. Precision takes into account that strange objects must be rejected.

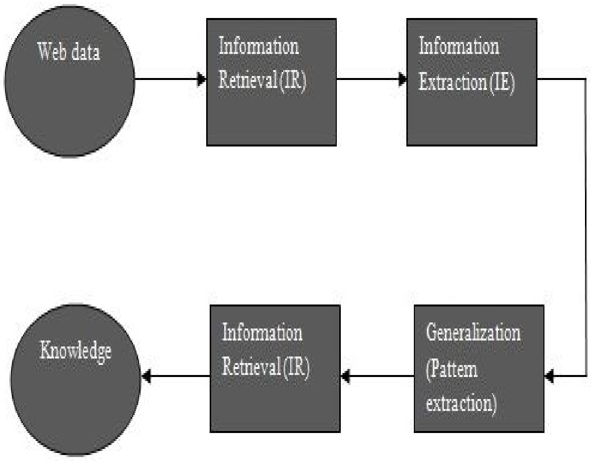

Web data
Information Retrieval (IR)
Information Extraction (IE)
Generalization (Pattern extraction)
Information Retrieval (IR)
Knowledge

Fig. 1. Mining Tasks

C. Information Extraction

Information Extraction (IE) is a process that extracts and retrieves information that is relevant to user based on the queries posted. Information extraction is the process of identifying main content of a web page or XML document which may consist of different forms of data in an unstructured and non-homogeneous manner. Added to this is the ability of including region and language based information, thanks to the exponential growth in use of cellular communication. Text based information has reached different levels with different languages forming the text either as a computer-generated data or acquired data through images forming most of the pages. All these aspects bring in a necessity of using a more general approach to extraction of information and it has become very important to consider different kinds of web pages or XML.

V. PROPOSED SYSTEM

A. Xsearch

In a keyword search, for a given keyword, there may be a many matches found in the data. Any of the matches can be or cannot be necessarily relevant to the query as expected. To solve this problem, many approaches are proposed to identify relevant keyword matches. XSearch is a semantic search engine for XML. It has simple query language, suitable for a naive user. As a result of search, it returns semantically related document fragments. XSearch method is used to search the semi structured data that present in the web pages or XML documents. It list all the text that relevant to the keyword. XSearch employs more information- retrieval techniques compared to XRANK. The XRANK used to ranking the documents. The highest ranking has the sufficient information relevant to the keyword. XSearch method retrieves the documents that has the attributes and relationships of entities and it can be modeled as graph..Extended information retrieval techniques are used to rank query answers. For efficient implementation, advanced indexing techniques are developed. It includes full-text search features and ranking to XQuery.

B. Artificial Neural Network

ANN consists of large number of simple processing elements that are interconnected with each other and layered also. NNs are designated by the network topology. In Neural network the entities are acts as entities. The network are modeled as entities and used to extract the entities, attributes of entities and their relationships from the semi structured data.

VII. EXPERIMENTAL SETUP

In order to evaluate the proposed method, four measured are compared on proposed approach with existing methods. Large number of XML document data set are available in web. In this paper, we examine a sample XML document. The sample XML document is

A. sample XML document

```
<bookstore>
<book category="COOKING">
 <title lang="en">Everyday Italian</title>
```

```
<author>Giada De Laurentiis</author>
<year>2005</year>
<price>30.00</price> </book>
 <book category="CHILDREN">
<title lang="en">Harry Potter</title>
<author>J K. Rowling</author>
<year>2005</year>
<price>29.99</price>
</book>
<book category="WEB">
 <title lang="en">Learning XML</title>
 <author>Erik T. Ray</author>
 <year>2003</year>
<price>39</price>
</book>
```

In this document, the keyword book can be used to search. Like this, the number of documents relevant to the keyword are retrieved. In this paper, we sample this XML document. Using XSearch technique, the keyword "Book" to search the relevant documents. The XRank method is ranking the documents which contains the rich contents. Then, Using Artificial Neural Network model, the book name as entity can be extracted from XML document. The attributes and relationship of entity book like category, author and price can be extracted and modeled as in a graph.

VIII. RESULT AND DISCUSSION

We measured the effectiveness of each system using precision, recall, F-measure, and accuracy. Precision measures the fraction of the documents correctly classified as relevant, while recall measures the fraction of relevant documents retrieved from the data set. Fmeasure is a single measure that tries to combine precision and recall. Accuracy measures simply the prediction correctness of the classifiers.

The experiment results on accuracy, precision, recall, and F-measure are summarized in Table I

Existing vs Proposed Method	Accuracy(%)	Precision (%)	Recall(%)	F- measure
SLCA	83.45	68.36	63.36	0.6577
MLCA	88.30	88.02	61.49	0.7240
XSEARCH(Proposed)	93.27	92.78	74.52	0.8265

In the above table, the accuracy denotes the ratio of number of documents correctly classified by the system to number of all documents in the tested set. Using this calculation, the XSearch method achieves the highest accuracy rather than SLCA and MLCA methods. The F- measure is based on the precision and recall. The F- measure is the ratio of product of precision and recall and the sum of precision and recall. Since F-measure represent a balance between precision and recall, the XSearch achieves highest accuracy and F- measure.

IX. COMPARISON CHART

The results can be compared with other methods and represent in a graphical format.

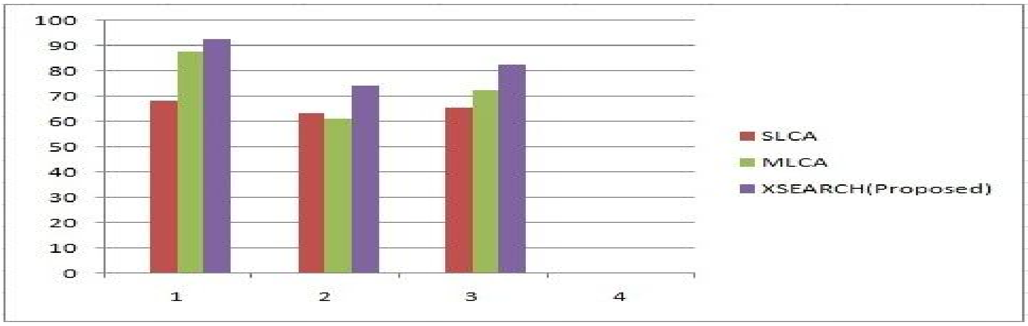


Fig. 2 : Accuracy Comparisons

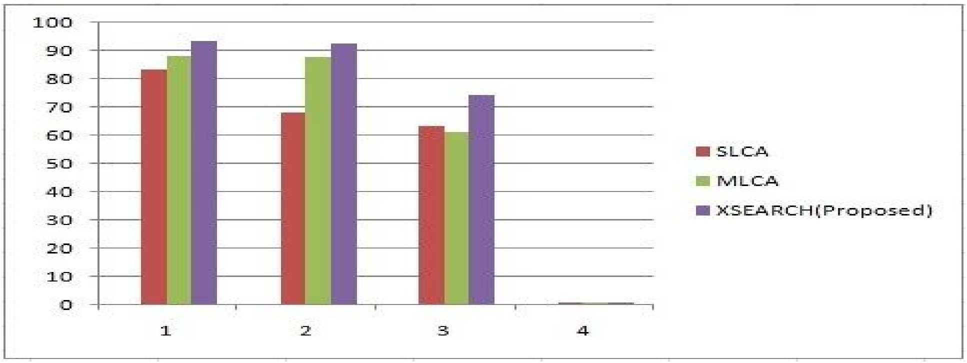


Fig.3: F- measure Comparisons

From the above charts from fig.2 and 3 the comparative analysis states the evaluation of proposed approach with an existing web content extraction method. The two classification algorithm describes information extraction to notice the web content without outliers

X. CONCLUSION

`In this paper the entities and their attributes from the semi structured information can be extracted. The semi structured text can be listed by using the XSearch keyword search method. After that, the Artificial Neural Network(ANN) model is used to extract the entities, attributes and their relationships from the semi structured information and can be modelled as graph. The XSearch method achieves the highest accuracy and F- measure.

REFERENCES

[1] Guo, L., Shao, F., Botev, C., Shanmugasundaram, J.: XRANK: ranked keyword search over XML documents. In: SIGMOD Conference, pp. 16–27 (2003)

[2] Ziyang Liu · Yi Chen , "Processing keyword search on XML: a survey", Springer Science+Business Media, LLC 2011. [3].HaiDong,Farookh Khadeer Hussain, Elizabeth Chang," A Survey in Traditional Information Retrieval Models", IEEE 2008

[3] Claire Cardie and David Pierce. Proposal for an interactive environment for information extraction. Technical Report TR98-1702, 2, 1998

[4] Rich Caruana, Paul Hodor, and John Rosenberg. High precision information extraction. In KDD-2000 Workshop on Text Mining, August 2000.

[5] L.R. Rabiner. A tutorial on hidden markov models. In Proc. of the IEEE, volume 77, pages 257–286, 1989.

[6] C. Chang, M. Kayed, M. Girgis, and K. Shaalan. A surveyof web information extraction systems. IEEE TKDE18(10):1411–1428, October 2006.

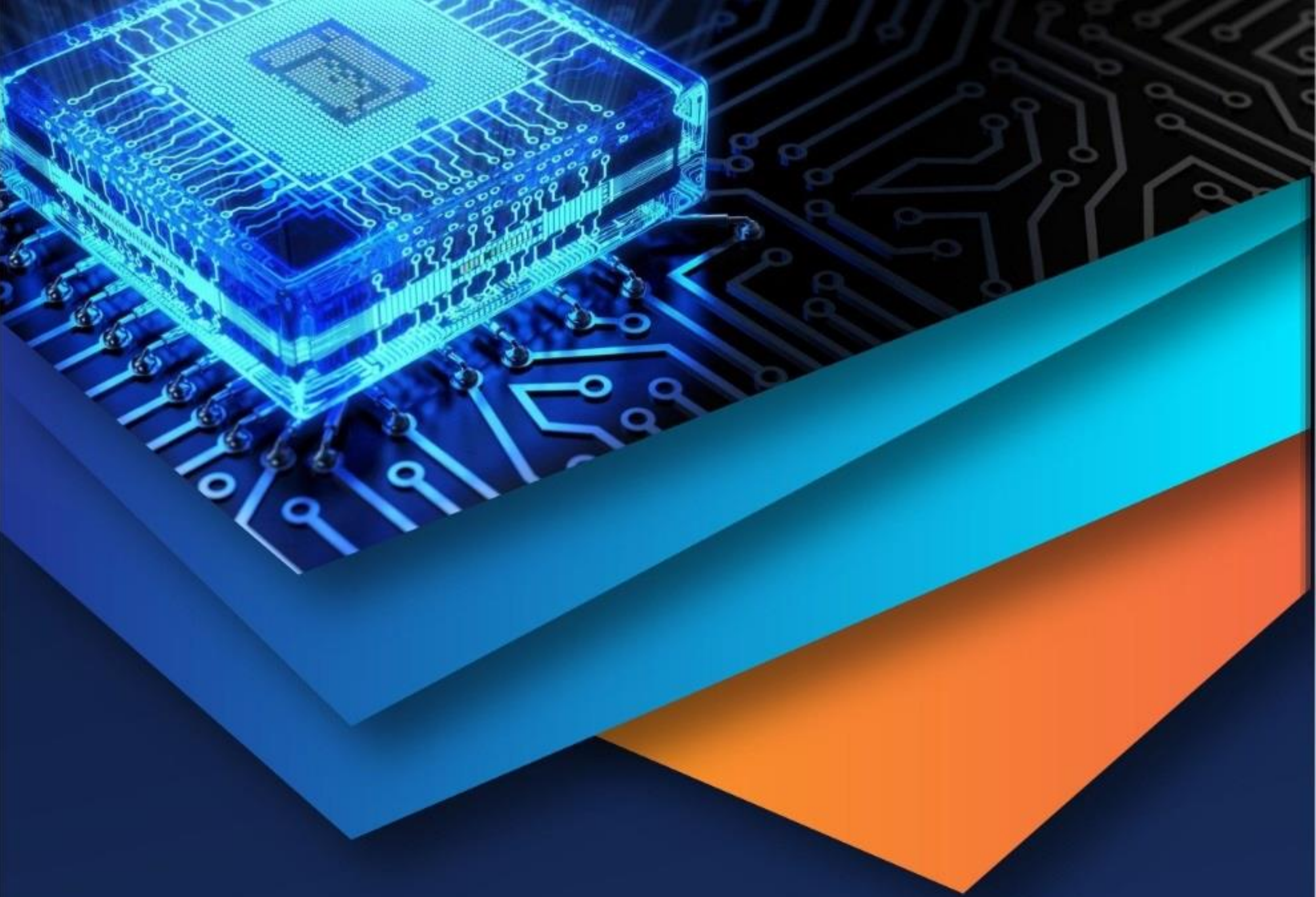
[7] J. H. Lim, "Visual keywords: From text retrieval to multimedia retrieval," in Soft Computing in Information Retrieval: Techniques and Applications, F. Crestani and G. Pasi, Eds. Heidelberg, Germany:Physica-Verlag, 2000, vol. 50, pp. 77–101.

[8] D. Merkl and A. Rauber, "Document classification with unsupervised artificial neural networks," in Soft Computing in Information Retrieval: Techniques and Applications, F. Crestani and G. Pasi, Eds. Heidelberg, Germany: Physica-Verlag, 2000, vol. 50, pp. 102–121.

[9] H. Fukuda, E. Passos, A. M. Pacheco, L. B. Neto, J. Valerio, V. J. D. Roberto, E. R. Antonio, and L. Chigener, "Web text mining using a hybrid system," in Proc. 6th Brazilian Symp. Neural Networks, 2000, pp. 131–136.

[10] H. Garis, C. Shuo, B. Goertzel, L. Ruiting, "A world survey of artificial brain projects, Part I, Largescale brain simulations", Neurocomputing, vol. 74, no. 1–3, pp. 3–29, August 2010.

[11] C. Lin and H. Chen, "An automatic indexing and neural network approach to concept retrieval and classification of multilingual (Chinese–English) documents," IEEE Trans. Syst., Man Cybern., vol. 26, no. 1, pp. 75–88, Feb. 1996.



10.22214/IJRASET

INDEX
COPERNICUS

45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY