



EXTRACTION AND SUMMARIZATION OF RELEVANT FORUM POSTS

Aboobacker Sidheeq. M
Computer Science and Engineering
Kathir College of Engineering
Neelambur, Coimbatore
sidheeqm@gmail.com

T.K.P. Rajagopal
Computer Science and Engineering
Kathir College of Engineering
Neelambur, Coimbatore

Abstract— A forum is a web application for holding discussions and posting user generated content. Messages within these forums or sub-forums are displayed either in chronological order or as threaded discussions. There are forums based on topics which acts like a big virtual information center. Information on wide range of topics is discussed through the forum. Here, we extract recent and relevant posts coming under a specific thread and summarizing it. Pattern matching and other data mining techniques are applied for extracting these forum posts. Existing Google indexing is utilized for extracting posts in threads. Posts are summarized by Latent Semantic Analysis and are further enhanced by Latent Dirichlet Allocation method.

Index Terms - Forums, Crawler, Latent Semantic Analysis

1. INTRODUCTION

Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. This enables much richer forms of queries on the abundant un-structured sources than what is possible with keyword searches alone. The extraction of structure [1] from noisy, unstructured sources is a challenging task that has engaged a veritable community of researchers for over two decades now. With roots in the Natural Language Processing (NLP) community, the topic of structure extraction now engages many different communities spanning machine learning, information retrieval, database, web, and document analysis. Early extraction tasks were concentrated around the identification of named entities, like people and company names and relationship among them from natural language text. Applications such as comparison shopping, and other automatic portal creation applications, lead to a frenzy of research and commercial activity on the topic. As society became more data oriented with easy online access to both structured and unstructured data, new applications of structure

extraction came around. Online communities are valuable information sources where knowledge is accumulated by interactions between people.

Forums are one of such communities which is a good source of information. Online forum pages are not the same as general web pages. My goal is to design effective retrieval models that incorporate online forum properties so that the effectiveness of forum search and hence posts extraction can be improved. The forums pages have unique textual or structural features that distinguish them from general web pages. Generally, a forum has several sub-forums covering high-level topic categories. Each sub-forum has many threads. A thread is a more focused topic-centric discussion unit and is composed of posts created by community members. This is an advantageous feature that encourages in-depth discussion, compared to general web pages.

Web forum has become an important resource on the Web due to its rich information contributed by millions of Internet users every day. In the forum sites people communicate and discuss with each other in a thread. A typical bulletin board is very popular almost all over the world for opening discussion. Every day, there are innumerable new posts created by millions of Internet users to talk about any conceivable topics and issues. Thus, forum data is actually a tremendous collection of human knowledge, and therefore is highly valuable. It is also noticed that some recent research efforts have tried to mine forum data to find out useful information such as business intelligence and expertise [2]. However, whatever the application is, the fundamental step is to fetch data.

In forum sites people can hold conversations in the form of posted messages. These messages are often longer than one line of text and are at least temporarily archived. Also, depending on the access level of a user or the forum set-up, a posted message might need to be approved by a moderator before it becomes visible. Depending on the forum's settings, users can be anonymous or have to register with the forum and



then subsequently log in in order to post messages. On most forums, users do not have to log in to read existing messages. Millions of posts are posting per day in the forum sites.

A post is a user-submitted message enclosed into a block containing the user's details and the date and time it was submitted. Posts have an internal limit usually measured in characters. Often one is required to have a message of minimum length of 10 characters. There is always an upper limit but it is rarely reached - most boards have 10,000, 20,000, 30,000, or 50,000 characters. Most forums keep track of a user's post count. The post count is a measurement of how many posts a certain user has made. Users with higher post counts are often considered more reputable than users with lower post counts, but not always. In this project I extract such posts related to the query and makes summary.

Summarization is the way expressing the ideas in fewer words. The introduced Summarization technique summarizes forum posts. This system extracts posts and provides a summary about the discussion. Two different techniques are used for summarization. Latent semantic analysis (LSA)[3] is a technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text. A matrix containing word counts per paragraph (rows represent unique words and columns represent each paragraph) is constructed from a large piece of text and a mathematical technique called singular value decomposition (SVD) is used to reduce the number of columns while preserving the similarity structure among rows. Words are then compared by taking the cosine of the angle between the two vectors formed by any two rows. Values close to 1 represent very similar words while values close to 0 represent very dissimilar words.

Latent Dirichlet Allocation (LDA)[4] is used to discover the topics that contained in the given set of sentences. LDA represents documents as mixture of topics that split out words with certain probabilities. Here, the individual posts are treated as documents and are having some fixed number of sentences and words. The summary produced by this generic model is more reduced than LSA.

2. RELATED WORK

iRobot[5] is one of the most important attempt to identify the Forum sites from their URLs. IRobot's first step is learning the sitemap of forum site. This is achieved by collecting pre-

sampled pages from the forum sites and then decides how to select an optimal traversal path to avoid duplicates and invalids. Forums sitemap is Re-constructing here. This paper builds a prototype of an intelligent forum crawler, iRobot, which has intelligence to understand the content and the structure of a forum site, and then decide how to choose traversal paths among different kinds of pages. After that, we select an optimal crawling path which only traverses informative pages and skips invalid and duplicate ones. One of the important advantages of this paper is the division of long threads into multiple pages and this can be reconstructed and concatenated, which is of great help for further indexing and data mining.

It is also noticed that some recent research efforts have tried to mine forum data [6] to find out useful information such as business intelligence and expertise. Forum pages are generated by pre-defined templates, and different templates are usually adopt to present different content such as list-of-board, list of-thread, post-of-thread, user profile, etc. Moreover, valuable information in forum pages is mostly shown in some repetitive manners, as it is essentially data records stored in a database. Thus, a forum page can be well characterized by what kinds of repetitive regions it contains, and where these regions are located. Location information should be integrated with the URL patterns to decide which links should be followed and how to follow them.

Jingtian Jiang designed and developed a supervised web-scale forum crawler, FoCUS[7]. The FoCUS aims to crawl relevant forum content with minimal overhead. This paper explains about the common pattern existing among forums and their threads even with different supporting software packages. FoCUS reduced the forum crawling problem into URL type recognition problem. Here scholars suggest some regular expressions as patterns for identifying URLs of forum threads. A single pattern can be used for crawling a large set of similar forum sites. Supervised method is applied here so the training set creation is tedious task. Special interest had been given by FoCUS to identify forums entry page by proposing an algorithm.

Similar work to crawl forum sites includes the work by Vidal et al. they proposed a method for learning regular expression patterns of URLs. This method compares DOM tree [8] structure and is effective for the sampled forum alone. Similarly FOCUS architecture has two major parts: Learning part to learn ITF regexes of a given forum from automatically constructed URL training examples. Second part is online crawling part to crawl all threads. The classifier used in



FOCUS is very effective. The method introduced by Koppula et al [9]. is used here to identify the different URL patterns and are refined recursively. This refined pattern is retained only if its matching URLs are greater than an empirically determined threshold. Online crawling is done by depth-first-strategy.

2.1 Feature Identification Methods

Summarization technique is employed to reduce the size of information. The system will summarize the posts and provide the user an overview about the posts. A Latent Semantic Analysis (LSA) based feature-identification approach works best to identify features. Features and opinion word identification are essential in feature-based summarization.

The effectiveness of Latent Dirichlet Allocation (LDA) for feature identification is examined in this paper.

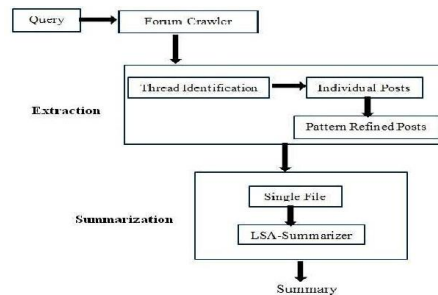
Latent Semantic Analysis (LSA) projects an original vector space or term-document matrix into a small factor space. The dimensional reduction of a matrix is accomplished using singular value decomposition which decomposes an original matrix into three matrixes, a document eigen vector matrix, an eigen value matrix, and a term eigen vector matrix. In turn, an original matrix can be approximated by multiplying these three matrixes with only high eigenvalues. Because of orthogonal characteristic of factors, words in a factor have little relations with words in other factors, but words in a factor have high relations with words in that factor.

Latent Dirichlet Allocation (LDA) incorporates the generative process of documents with Dirichlet distribution. According to LDA process, each document is generated in the following three steps. First, the number of words used in a document is determined by sampling with the Poisson distribution. Second, a distribution over topics for a document is elicited from the Dirichlet distribution. Third, based on the document-specific distribution, topics are generated, and then words for each topic are generated. LDA also provides topics in which words have probability values. Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

3. PROPOSED SYSTEM

A detailed description on the implementation aspects of the proposed work is discussed in this chapter.

3.1 System Architecture



3.2 System Modules

The collection of valuable user generated posts and their summary generation process is done by dividing the task into four different modules. These modules are implemented[10] in the order as they are given in the following section. Details of this implementation process is given in next section of this chapter.

3.2.1 Document collection Module

Before implementing A Forum crawler some forum sites are deeply studied. Based on these studies , document collection module was started;[11] This step includes finding the best forums and threads for the posts that is required. Forums selection is done after fixing the domain of the discussions carrying out. Most of the data mining projects utilize the data warehouse to collect the required data. Here we need to extract from online forums and we need online analytical processing to collect required data. For this purpose pattern matching technique is used. Pattern is written in regular expression and is made possible by continuous analysis of the forum sites.

3.2.2 Document collection Module

The extracted posts [12] are selectively discarded before making their contents as a single file for making summary. Discarding posts by the forum etiquettes and by considering the sentence. All these elimination is done by constructing regular expression corresponding to each rules [13] in the forum etiquette.

3.2.3 Information Extraction Module

The posts are extracted and saved in temporary file [14] because the queries are different each time. This temporary file is used by the summarization Module.

3.2.4 Summarization Module

Fuzzy logic is a form of knowledge representation suitable for



notions that cannot be defined precisely, but which depend upon their contexts. It deals with reasoning that is approximate rather than fixed and exact. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. It can also use with linguistics. Linguistics is the scientific study of language. There are broadly three aspects to the study, which include language form, language meaning, and language in context. Fuzzy logic provides an alternative way to represent linguistic and subjective attributes of the real world in computing. Here for the summarization purpose I considered the two existing methods Latent Semantic Analysis and Latent Dirichlet Allocation.

4. IMPLIMENTATION DETAILS

All experiments were performed on a Windows 8 machine with two 2.3 GHz core i3 processors and 2GB memory. Forums are well structured and are organized as Index page, Thread Page and post pages respectively.

4.1 Extraction of Relevant Posts

In order to extract the forum posts it is necessary to know the depth of forums and the amount of contents, the user posts, in them. It is difficult and not necessary to crawl all the posts. Few of the forums are analyzed before implementing a post extractor. This has been generalized to fix the number of relevant posts to be extracted from forums. The basic forum structure includes three different types of pages they are the following. Board page is the index page of the Forum site it consist of some sub forums which is specific to the domain of the Forums. This is looking like:

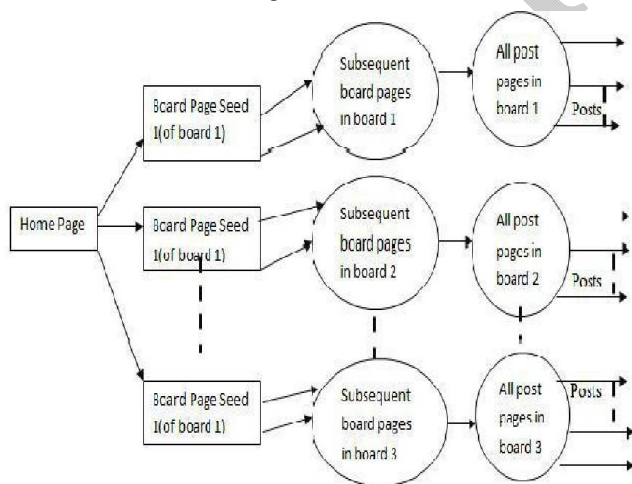


Figure 4.1: Basic Forum Structure

Main Forums	Last Post	Threads	Posts
Supplements Everything about the world of supplements!	Has anyone tried Follidrone by REALPistonFlorida Today, 05:56 PM	641,620	10,234,502
BodySpace Profiles And BodyBlogs View other BodySpaces and start your own!			
Workout Programs What workout is best for your goal?	Jakked and Tan with your... by PAC-mind9 Today, 05:56 PM	247,729	4,229,513
Exercises Post questions about specific exercises here.	I really need help with my... by YolkesPinky Today, 05:50 PM	193,433	2,453,996
Nutrition Everything related to proper nutrition.	Anyone else feel amazing... by KettLove Today, 05:56 PM	313,273	3,774,197

Figure 4.2: Board Forum Page

Title / Thread Starter	Last Post By	Replies	Views
Sticky: ATTENTION NCAA ATHLETES: Banned Substance Thread 440kiloj, 01-31-2012 11:07 AM	Sys11 06-04-2014, 05:11 PM	350	104,339
Sticky: Beginner's Creatine, Protein & Misc. Supplement FAQ *Must Read* DASLURKER, 03-07-2014 06:45 PM	marketing1 06-04-2014, 02:36 PM	52	7,670
Sticky: Bodybuilding.com warehouse pics ctopus, 04-09-2009 01:07 PM	zohBOT 05-29-2014, 02:30 AM	675	369,927
Sticky: -- The Idiot's Guide To Using the Search Bar -- Screenshots intended slapp, 07-31-2011 08:12 AM	DavidNatural 05-29-2014, 02:00 AM	184	107,615
Sticky: Overall Health-Oriented Supplements user27629377, 01-15-2009 08:18 AM	quill2214 05-25-2014, 09:59 AM	800	255,259
Sticky: General Forum Rules + Company Rep Rules ForumSentinel, 12-05-2012 09:22 AM	ForumSentinel 05-19-2014, 08:32 AM	2	463

Figure 4.3: Thread Page

Post page holds the posts or opinions of members temporarily. Each page is having a capacity of keeping around 30 posts. This can be viewed as:



Figure 4.4: Post Page

The proposed system accepts questions or issues from the user as the search query. These queries are related to different domains. The same domain has different topics and so has to face many queries. Forums are generally domain specific and have sub forums to be more specific in that domain. Structure of forums is different but in specific format, there will have a lot of forums. This makes the task of creating template for specific package and structure. This template is used for identifying the thread and through that relevant posts are extracted. Output of the extraction phase is the posts related to the search query. The query is compared with thread headings in the forum sites and most appropriate and relevant posts are extracted with the help of a good regular expression. This system utilizes the existing index of the Google [15] and string matching algorithm used in this. The requirement for proposed system is now available. To reach the defined goal it is necessary to follow the below phases.

4.2 Feature Based Summarization



Feature identification algorithms are employed to find out related feature terms of specified seed features, and these related terms could be regarded as being semantically related to the specified features. These related terms can be employed to select summary sentences. The two feature identification algorithms examined in this work are described below:

Algorithm 1: LSA for feature Identification

Input:

A $n \times m$ term-document matrix M , feature seed set S , n number of extracted features for each seed f in S

Output:

An association array F , where each key represents a feature seed f and its corresponding value is f 's related features

Steps:

1. Initialize associated array F .
2. Call `SingularValueDecomposition(M)`.
3. Assign the product of term eigen vector and eigen value to M .
4. For each f in S :
5. Assign `getTermVectorFromTermDocMatrix(f,M)` to wf .
6. Initialise similarity list sim .
7. For each column vector w of M :
8. Assign $wf : w$ to $sim[i]$.
9. Sort(sim).
10. Assign `getTopRelatedFeatures(sim,n)` to F .
11. Return F .

In algorithm 1, line 2 calls the Singular Value Decomposition function. It decomposes the M matrix into its corresponding term eigen vector, eigen values and document eigen vector. Line 3 assigns the product of term eigen vector and eigen value to N matrix. In lines 4-8, for each seed feature get the term vector wf from term document matrix and compute its dot product with each element in the column vector of N .

Then sort them in increasing order in order to get the top related features.

Algorithm 2: LDA for feature identification

Input:

A set of documents, fixed number of K topics, n number of extracted features for each seed topic f of K

Output:

An association array F , where each key represents a topic f of K and its corresponding value is f 's related features :

Steps:

1. Initialize associated array F .
2. Initialize each word in the document to one of the K topics.
3. For each document d , repeat until no change:
4. For each word w in d :
5. For each topic t :
6. Compute $p(\text{topic } t \mid \text{document } d)$.
7. Compute $p(\text{word } w \mid \text{topic } t)$.
8. Reassign w a new topic, where we choose topic t with probability $p(\text{topic } t \mid \text{document } d) * p(\text{word } w \mid \text{topic } t)$.
9. Update the n words assigned to F under each aspect.
10. Return F .

In algorithm 2, line 2 randomly assigns each word in the document to one of the K topics. Lines 3-7 computes the product of the proportion of words assigned to a topic in each document and the proportion of assignments to a topic throughout all the documents coming from this word w . Line 8 assigns w a new topic based on the repetition of the previous steps, until no change in the above computed value. Line 9 updates assignment of words under each topic.

5. RESULTS AND OBSERVATIONS

Results of the two important phases of this thesis work can be given as:

5.1 Posts Extraction

Many different measures for evaluating the performance of information retrieval [16] systems have been proposed. The measures require a collection of documents and a query. All common measures described here assume a ground truth notion of relevancy: every document is known to be either relevant or non-relevant to a particular query. In practice queries may be ill-posed and there may be different shades of relevancy. This chapter discusses about the results obtained from the proposed system. The proposed system is an attempt to implement information extraction from discussion forums.

Data set contains n number of posts downloaded from discussion thread.

To evaluate this system, Precision and Recall are used as evaluation metrics. Precision- is the fraction of the documents retrieved that are relevant to the user's information need. Recall-is the fraction of the documents that are relevant to the query that are successfully retrieved. We provide the Precision of downloaded posts and the Recall of posts to evaluate the effectiveness of this crawler. We suppose that, for crawling a Web forum site, Posts-downloaded is the count of all posts that could be download, and Posts-effective is the count of all



post that are chosen in summary generation, and Post-all is the count of all posts in the related topic. Then the Precision and Recall are computed as follows:

$$\text{Precision} = \text{Post-effective} / \text{Posts-downloaded} * 100$$

$$\text{Recall} = \text{Post-effective} / \text{Post-all} * 100$$

Table 5.1: Precision table

5.2 Summarization

Two different feature identification algorithms, namely LSA and LDA are compared [4]. The results of each of which are then used in generating the summary. Precision, recall and f-value are the matrices used for reporting the performance of feature identification algorithms. Precision is the ratio of the number of correctly extracted features to the total number of features extracted. Recall is the number of correctly extracted features to the total number of standard features. F-value gives the ratio of twice the product of precision and recall to their sum.

Precision values are higher for LSA and as the number of terms increases a significant improvement can be seen for LDA based feature identification. As far as recall is concerned, again LSA shows better results when compared with LDA but its performance begins to degrade when the number of terms increases and LDA seems to show good recall values at this point. Therefore LSA based feature identification gives better results than LDA based feature identification.

6. CONCLUSION AND FUTURE WORK

In this paper, we design and develop an intelligent Forum crawler. This Forum crawler extracts posts from relevant discussion threads of the forum. Pattern matching and data mining techniques are applied for extracting the forum posts. Before extracting the posts into separate files for summarization, specific rules are applied to remove certain words, sentences or posts. This makes posts more effective for applying machine learning algorithms and natural language processing technique. As the forums are rich in posts or relevant contents this technique should be enhanced by extracting grammatically correct sentences or posts. Here, these extracted posts are summarized by Latent Semantic Analysis and enhanced by Latent Dirichlet Allocation method. It would be interesting to extract high quality posts for decision making. Extract grammatically correct posts and give score to users for understanding their value in decision making. Forum users were unaware of the forum etiquettes and importance of their posts. This area also needs improvement. Moderators of most of the forum sites are inefficient so enhancement in this side is also necessary. Pattern matching by using regular expression is used in this

system to extract the posts. An obvious future work would be using natural language processing for extracting large training data. This system is domain and structure specific. Another future work is Scaling the system to a generic domain. The summary generation for the posts will be more accurate if we make the above enhancement. The behavior of forum members can be predicted by enhancing this System.

REFERENCES

- [1] A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in proceedings of the 2003 ACM SIGMOD international conference on management
- [2] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo, "Deriving marketing intelligence from online discussion," in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005, pp. 419-428.
- [3] S. T. Dumais, "Latent semantic analysis," Annual review of information science and technology, vol. 38, no. 1, pp. 188-230, 2004.
- [4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022, 2003.
- [5] R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, "iRobot: An intelligent crawler for web forums," in Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp. 447-456.
- [6] Y. Guo, K. Li, K. Zhang, and G. Zhang, "Board forum crawling: a web crawling method for web forum," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence. IEEE Computer Society, 2006, pp. 745-748.
- [7] J. Jiang, X. Song, N. Yu, and C.-Y. Lin, "Focus: learning to crawl web forums," Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 6, pp. 1293-1306, 2013.
- [8] Y. Zhai and B. Liu, "Structured data extraction from the web based on partial tree alignment," Knowledge and Data Engineering, IEEE Transactions on, vol. 18, no. 12, pp. 1614-1628, 2006.
- [9] H. S. Koppula, K. P. Leela, A. Agarwal, K. P. Chitrapura, S. Garg, and A. Sas-turkar, "Learning url patterns for webpage de-duplication," in Proceedings of the hird ACM international conference on Web search and data mining. ACM, 2010, pp. 381-390.
- [10] G. Gupta, Web Data Mining, in Introduction to Data Mining with Case Studies. New Delhi, India: Prentice-Hall, 2006.
- [11] Forum Matrix," [Online] Available :<http://www.forummatrix.org/index.php>, accessed Feb 2014.
- [12] Cordeiro and P. Brazdil, "Learning text extraction rules, without ignoring stop words." in PRIS. Citeseer, 2004, pp. 128-138.
- [13] C.-H. Chang, C.-N. Hsu, and S.-C. Lui, "Automatic



information extraction from semi-structured web pages by pattern discovery," Decision Support Systems, vol. 35, no. 1, pp. 129-147, 2003.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction for the web," in IJCAI, vol. 7, 2007, pp. 2670-2676.

[15] S. Brin and L. Page, "The anatomy of a large-scale hyper textual web search engine," Computer networks and ISDN systems, vol. 30, no. 1, pp. 107-117, 1998.

[16] M. Hu and B. Liu, "Mining and summarizing customer reviews," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004, pp. 168-177.

AUTHORS BIOGRAPHY

[Sciences](#) in WASET.



Aboobacker Sidheeq .M:

He Completed UG from Sri Jayaram Engineering College, Cuddalore, Tamil Nadu, India. He is Currently doing PG in Computer Science and Engineering from Kathir College of Engineering, Coimbatore, India. He is an active member of IEEE student branch. Attended various

International, National Conferences and Attended various work shop based on Image processing and Data mining . His research interests are Image processing, Data mining.



T.K.P. Rajagopal M.A., M.Phil., M.C.A., M.Phil., M.E.(CSE), M.Tech.(IT), Ph.D', is working as an Associate Professor in the Department of Computer Science and Engineering at Kathir College of Engineering, Coimbatore, Tamailnadu, India. He has

15 years of teaching experience. His research areas are Network Security, Data Mining and Digital Image Processing. He has published 3 books, 9 research papers in international journals and 2 research papers in national journals, Presented 17 papers in International Conference, 19 papers in National level conferences and has attended 3 QIP, 5 FTP, 10 Seminars and 33 Workshops. He is a life member of various professional societies like ISTE, DOEACC, CSTA, CSI, ACS, IAS, IACSIT, IIRJC, SDIWC, ISI, IoN, UACEE, theIREDD, IAENG Society of Artificial Intelligence, Computer Science, Data Mining, Software Engineering, and Wireless Networks, Hong Kong, Serving as a reviewer in IJITCE, WCSIT, CSC-IJCSS, CSC-IJCN, IJCSIS, [Scientific & Technical Committee & Editorial Review Board on Engineering and Natural](#)