IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol-6, Special Issue-3, Apr 2018 E-ISSN: 2347-2693

Extractive Approaches for Automatic Text Summarization

S. Gandotra[1*], B. Arora[2]

[1*]Dept. of Computer Science & I.T, School of Applied Sciences, Central University of Jammu, Jammu, India
[2] Dept. of Computer Science & I.T, School of Applied Sciences, Central University of Jammu, Jammu, India

[*]*Corresponding Author: sonam2gandotra@gmail.com, Tel.: +91-9796699466*

Available online at: www.ijcseonline.org

Abstract— With the flooding of a huge amount of data on the web with reference to the text, a technique to condense this data in summary form is very important so that the users can have access to relevant information regardless of enormous content on the web that is available to the user. This content could be informative, relevant or even important to the user or could be purely irrelevant. Text summarization techniques help in reducing the time and effort of the user looking for content about a particular topic on the internet by summarizing the content of the documents and the user by only looking at the summarized content can decide whether the document is relevant or irrelevant. Thus, automatic text summarization techniques play a key role in information retrieval from the web. In this paper, a study of various text summarization techniques has been conducted based on parameters like a number of documents, content, output, language, availability of training data etc. Also, the summary evaluation processes i.e. intrinsic and extrinsic are discussed. Extractive approaches for text summarization are also discussed and the recent work done in each of these approaches is compared and contrasted.

Keywords— Text Summarization, Extractive summarization, Intrinsic Evaluation, Extrinsic Evaluation

I. INTRODUCTION

Automatic text summarization is the process of condensing a large amount of text into a precise summary without affecting its actual meaning. Although, its roots can be found in the late 50's [1] but with the increase in the online addition of content on the web and automation of systems, its need has emerged even more. With the huge bundle of information present on the web about even a small topic, it becomes difficult for the user to extract relevant information and most of the time's user ends up getting irrelevant information. This whole process of information retrieval can be very cumbersome in absence of tools like automatic summarization and data mining. Automatic summarization has highly improved the information extraction process, the popular search engines like Google, yahoo and Bing etc. also rely on this technique for better information retrieval and providing a better experience to the users. A robust system which can compress information from various documents into a shorter readable summary is the need of the present generation. As humans, we can easily summarize the text, images or videos by going through them but for the machine to read this kind of data and interpret it so as to maintain the core of the given content is a very challenging task. Efficient natural language processing, linguistic and statistical methods have to be applied so as to get the desired result. Hence, it is considered as the most challenging task in NLP as coverage, cohesion, redundancy and significance of information has to be taken care of[2].

With time, various techniques have been proposed for summary generation each fulfilling a particular criterion .These techniques can be classified on various parameters like dimension, output, language, learning, context and information as shown in table 1.

Table 1. Classification of Summarization Techniques

Parameter	**Summarization technique**	
Dimension	Single- document summarization	Multi-document Summarization
Output	Abstractive Summarization	Extractive Summarization
Language	Mono-lingual Summarization	Multi-lingual/Cross-lingual Summarization
Learning	Un-supervised Summarization	Supervised Summarization
Context	Generic summarization	Query-based Summarization
Information	Indicative Summarization	Informative Summarization

Summary evaluation is also one of the important tasks in text summarization to know the validity of the summary being generated. Evaluation generally depend on two parameters:

Compression ratio(how shorter the generated summary is) and retention ratio/omission ratio(how much information the summary contains)[3]. Intrinsic and extrinsic evaluations are used to determine the performance of the summary based on above two parameters. Intrinsic evaluation involves comparing the gold (generally human generated) summary with the automatic summary generated. The main focus is on informativeness and sentence cohesion. The measures used for intrinsic evaluation are shown in Figure 1.

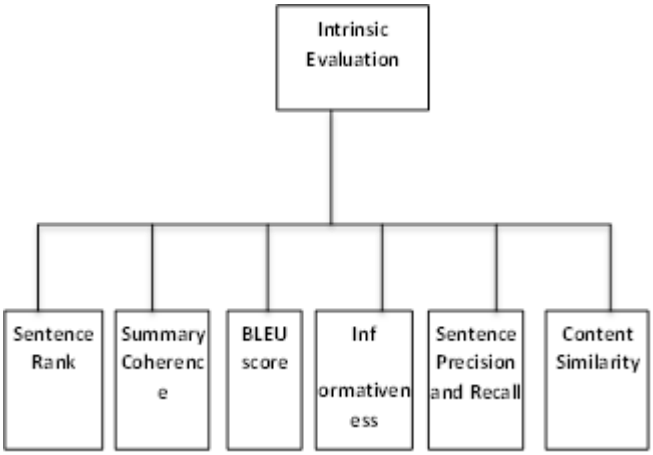

Figure 1. Measures of Intrinsic Evaluation[3]

Extrinsic evaluation measures the acceptability and efficiency of the summary generated for other tasks like information retrieval system, question answering system etc. The two common techniques used in this evaluation are reading comprehensions and relevance assessment. The measures used for extrinsic evaluation are shown in Figure 2.

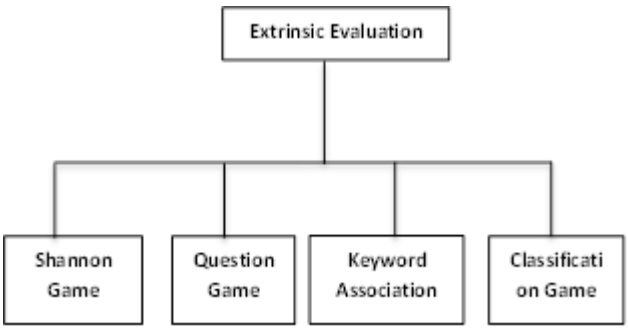

Figure 2. Measures of Extrinsic Evaluation[3]

The paper is organized in four sections. Section-I gives the introduction about the idea of automatic text summarization while in section-II extractive summarization is discussed in detail. Extractive approaches for summarization are defined in section-III along with comparative analysis of the work done in all four approaches i.e. statistical, topic-modeling, graph-based and machine learning techniques. Finally, conclusion and future scope is presented in section-IV.

II. EXTRACTIVE SUMMARIZATION

The process of selecting a set of sentences from the given document which represent the whole idea of the document is known as extractive summarization. Sentences are chosen as they are written in the text to generate the summary depending upon the compression rate provided by the user. Key features are extracted from the document and then sentences are ranked according to the value of key features. Extractive Summarization techniques mainly rely on the fulfillment of the following criteria:

- *Coherency*: To check whether the extracted summary is in proper order or not.
- *Coverage*: To determine whether the final summary represents the whole aspects of the document or not.
- *Redundancy:* To check for redundancy in the final summary. It occurs mainly in multi-document summarization.
- *Significance*: To check whether the summary is as informative as the document or lacks the correct description of the document.

The concept of extractive summarization is defined in figure 3.

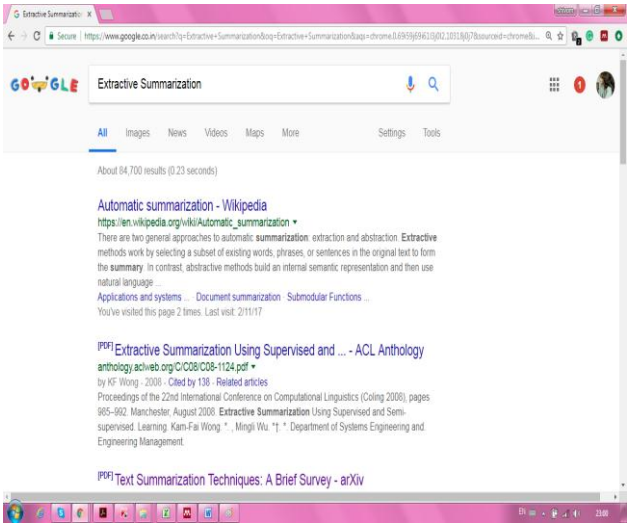

Figure 3 Extractive Summarization[4]

The figure shows how Google search engine uses extractive summarization to make its results more relevant to the user. The content below the main heading represents the extractive

summary of the corresponding documents. Thus, summarization techniques help in getting the gist of the whole document on the web and thereby, making it easy for the users to decide whether they want to go through the whole document or not.

III. EXTRACTIVE APPROACHES FOR SUMMARY GENERATION

A. Statistical Based Approach

These are language- independent statistical approach used to generate summary of the given text. These on the basis of extracted key feature of the document assign score to the features and then the highest score features are used to create the summary. Various factors which are considered are: position of sentence, keywords(positive/negative), relative length of sentence, resemblance to title, presence of numeric data, presence of named entities in the document, information gain, centrality of sentence, term frequency-inverse document frequency(TF-IDF), presence of cue words. All or some of these features have been considered by the researchers for extracting the summary of the document.

Multi document summarization based on single document summary cluster was given by author in paper[5]. First preprocessing of documents is done by removing the stop words, stemming, tokenization, frequency computation and splitting. Then the statistical features i.e. location, theme similarity, reference index of sentences, document feature are extracted from each document. Based on compression ratio, summary of each document is generated by using sentence weight. Syntactic and semantic similarity measures as proposed by [6] are used for calculating similarity of sentences. Shortest path length, depth of subsumer and information content are used for finding the semantic similarity between words. To form the multi document summary, clustering of sentences is done using the sentence similarity measure. The proposed model is tested on DUC 2002 dataset and the evaluation of the summary is done using ROGUE score.

A sentence scoring assessment technique is proposed by the author in paper[7]. Both qualitative and quantitative assessment is done for the proposed model. Scoring is done on 15 parameters which are broadly divided into 3 categories: word score(word frequency, word co-occurrence, proper noun, upper case, tf-idf, lexical similarity), sentence score(cue phrases, length, inclusion of numerical data, resemblance to title, centrality, position) and graph score(aggregate similarity, text rank, bushy path of node). Three different datasets are used for evaluating the performance of these scores i.e. CNN dataset, blog summarization dataset and SUMMAC dataset. Quantitative assessment is done using the ROUGE parameter while for qualitative assessment human experts are employed.

The author in paper [8] devised a new technique for text summarization of Hindi documents by focusing on 11 features of documents as Term Frequency-Inverse Sentence, Length of the sentence in document , Location of sentence in the document, Similarity between sentences, Numerical Data, Title Overlap, Subject Object Verb (SOV) Qualifier, Subject Similarity , Hindi Cue Phrase Feature , Common Hindi-English , Presence of Email Addresses and after structured representation of the document, it applies fuzzy logic to the data and obtain the summary of the given documents. The documents taken are the news articles from different online news channels about the same article.

A multi-document summarization framework has been developed by author in paper [9] based on sentence cluster using NMTF(Nonnegative Matrix Tri-Factorization). Both the inter type relation and intra type relations are used to generate the summary based on sentence, document and term parameters. Then the performance of the given framework is evaluated on the DUC-2004 and TAC-2008 datasets with the existing frameworks for summarization with only inter-type relations.

The analysis of the above mentioned literature is presented in table 2. It shows various parameters being taken up by the researchers to formulate summary.

Table 2. Analysis of Statistical-based Approach

Parameter / Paper	Gupta et.al [5]	R. Ferreira et al. [7]	Gulati et.al [8]	Yang et.al [9]
Positional feature	✓	✓	✓	✓
Frequency Count	✓	✓	✓	✓
Sentence Centrality	✓	✓		✓
Resemblance to title		✓	✓	✓
Relative length	✓	✓	✓	
Cue-phrases		✓	✓	
Proper Noun		✓	✓	
Numerical Data		✓	✓	
Bushy Path		✓		
Aggregated Similarity		✓	✓	✓
Term Frequency-Inverse Document Frequency(TF-IDF)	✓	✓	✓	✓
Information Gain				✓
Lexical Similarity	✓	✓	✓	✓

Text Rank		✓		✓
Word Co-occurrence		✓		✓
Presence of e-mail			✓	
Language	English	English	Hindi	English
Document-type	Multi-document	Single-document	Multi-document	Muti-document
Dataset	DUC-2002	CNN, SUMMAC, Blog Summarization	News articles from online news channel	DUC-2004 and TAC-2008
Evaluation Parameter	ROUGE	ROUGE and Human Experts	Precision, Recall and F-measure(ROUGE)	ROUGE-1, ROUGE-2 and ROUGE-SU4

B. Topic- based approaches

In topic-based approach rather than using the sentence-term factors into consideration, the topics embedded in the documents are recognized. This approach has two main advantages: the context of the document becomes clear which is absent in statistical approaches and different themes or topics present in the document are also identified which again add to the weight of the summary generated.

A contextual topic modeling approach based on the hierarchal Bayesian model for multi-document summarization is proposed by author in paper [10]. Distinction of general topics from the specific topics is done and their correlation is also calculated which helps in analyzing the topic hierarchies. Query focused summary is produced by taking into account the relevant sentences. The model was tested on the TREC and AQUINT dataset of DUC-2005 and DUC-2006 respectively which are then evaluated using the ROUGE parameter.

An extractive approach for novel summarization based on topic modeling is proposed by author in [11]. Diversity of topic, its distribution in the whole text and compression ratio are the fundamental aspects of the approach as proposed by the author. First preprocessing is done to remove noisy elements from the text, then topic modeling based on LDA algorithm (Latent Dirichlet Allocation) is applied which based on the probability distribution of each topic traces the sentences associated with it to form the candidate set of sentences. In the third step, based on sentence position, sentence importance, positive/negative diversity, redundancy rate sentence are selected from the candidate set to form the summary. Finally smoothing is done by applying external sources like, thesaurus and synonyms to improve the machine readability and thus resulting in a better quality of summary.

A brief analysis of the above mentioned text is shown in table 3.

Table 3. Analysis of Topic-Modeling based approach

Parameter / **Paper**	**Yang et.al** [10]	**Wu et.al** [11]
Document Type	Multi document	Single document
Summary type	Query-based	Generic
Algorithm used	Hierarchal Bayesian	LDA(Latent Dirichlet Allocation)
Dataset	Part of DUC-2005 and DUC-2006	63 narrative summaries as the novel dataset
Evaluation Parameter	Perplexity, ROUGE-1,2,SU4	Manual approach and rouge parameter
Focus	To convey the contextual information for determining similarities between different documents	Topic diversity and compression ratio

C. Graph-based approaches

In graph based approach, the sentences or terms in the document are considered as nodes and the relationship between these sentences is represented as edges connecting the nodes. The importance of a sentence is measured by the no of connection it has with other nodes. This approach mainly relies on the sentence centrality and semantic similarity between inter and intra elements of a given dataset for single and multi-document summarization respectively.

GraphSum, a graph based approach using correlation mining for multi-document summarization was introduced by author in paper [12]. The data is first preprocessed by stop word removal and lemmatization to keep the algorithm language independent. After preprocessing, association rule mining is applied to find the correlation graph of the given documents which lays more importance to frequent item set association and positive and negative correlations to know the proper context of relations between the sentences. Then indexing is

performed by using a variant of the PageRank algorithm to find the relevance of the extracted data. Sentence relevance score and sentence coverage are used as the essential parameters for selection of sentences from the correlation graph. The validity of the above algorithm is check on the DUC-2004 dataset and five news article collection of the leading newspapers. ROUGE is used for evaluating the performance of the system. The results produced are at par with state-of-the-art algorithms.

A multi-graph based approach has been introduced by the author in [13]. This approach relies on the relationship between sentences rather than words to avoid ambiguity. Also the no. of edges is equal to the no. of common words appearing in sentences. After the construction of graph, a matrix is constructed which highly reduces the dimension of summary yet containing the main idea of the article. Also, cosine similarity which is mostly used by other researchers to find the sentence similarity is not used here. Proposed approach is tested on a set of more than 1000 passages which are further divided into three datasets and is also evaluated on the ROUGE parameter.

CoRank, a single document extractive summarization approach which focuses on word-sentence relationship and presented this relationship in the form of graph is presented by the author in paper [14]. Graph based ranking model is used to incorporate the relationship of words with sentences to more accurately understand the context of the document. Also, redundancy control is also proposed which can further enhance the quality of the summary produced. DUC-2002 and 10 Chinese documents are used for testing the relevance of the approach. Recall, Precision and F-measure are used for evaluating the results with the other earlier approaches.

An analysis of the above discussed approaches is given in table 4.

Table 4. Analysis of Graph-based approach

Parameter / **Paper**	**Baralis et.al**[12]	**Fatima et.al**[13]	**Fang et.al**[14]
Document Type	Multi-document	Single-document	Single-document
Language	English	English	Chinese, English
Dataset	DUC-2004 and 5 news article collection	1,000 text passages divided into 3 dataset	DUC-2002 and 10 Chinese documents
Model used	Correlation graph based on association rule mining	Multi-graph based model	Word-sentence correlation presented in graph based model
Evaluation Parameter	ROUGE-2, ROUGE-SU4	Recall, Precision and F-score	F-score, ROUGE-1,2,L

D. Machine learning based approaches

Machine learning approaches depend on the availability of training data from which they learn how to act according to the given input parameters. These approaches can either be supervised, unsupervised or semi-supervised depending on availability of training data. In text summarization, mostly supervised and unsupervised approaches are considered. In supervised learning approach, the text along with human generated summary is fed for training the data and then by identifying relationship between the text and summary, further summaries are generated. Given text is classified into two regions: one belonging to final summary and other not belonging to final summary. Various techniques like Naïve Bayesian Classification, Support Vector Machines (SVM), Regression, Decision Tress etc. are used in supervised learning.

In unsupervised approach, no training data is fed into the system and system automatically clusters the given set of data into different clusters. From these clusters, sentences of utter importance are extracted to form the final summary. Various techniques like K-mean clustering, nearest neighbour, Hidden Markov Model (HMM) etc. are used in unsupervised learning. In semi supervised learning, both the labeled and unlabeled data is used for training the given input.

Genetic Algorithm (GA), Mathematical Regression (MR), Feed Forward Neural Network (FFNN), Probabilistic Neural Network (PNN) and Gaussian Mixture Model (GMM) are trained on statistical features using 100 and 50 manually generated summary articles in Arabic language and English language respectively by the author in [15]. After training the models on these algorithms, the models are tested on 100 Arabic and English documents. The algorithms are trained on various feature set like sentence centrality, negative/positive keywords, resemblance to title, aggregate similarity etc. Some of the features are language dependent while others are language independent. Sentences are then ranked according

to the value of features as calculated by the corresponding algorithm and the highest ranked sentences are used for summary generation depending on the compression ratio as defined by the user. Out of all the algorithms tested, GMM gives the best result.

A sentence clustering framework based on ranking for multi-document summarization is proposed by the author in paper [16]. Two ranking functions i.e. simple ranking and authority ranking are used for calculating the conditional rank of terms and document in the given dataset. Then ranking based clustering is used to find the similarity between the clusters rather than calculating the similarity between terms, sentences or documents. Spectral approach [17] is used for predicting the expected no. of clusters in the dataset. Highest ranked sentences from the theme cluster which is ranked highest is used for extracting summary sentences. The framework is evaluated on both intrinsic and extrinsic parameters using DUC-2004 and DUC-2007 datasets.

A supervised machine learning based approach using Support Vector Machine (SVM) for Hindi language is proposed by author in paper [18]. SVMs are used for training the model based on the value of feature set. The feature set contains parameters like numerical data, sentence position, length of sentence, keywords etc. The algorithm classifies the sentences into four rank categories i.e from 1 to 4 representing least important to most important sentences respectively. Based on the user defined compression ratio, the sentences are extracted from the document. Hindi news document from leading newspaper are used for testing the proposed framework.

A brief analysis of the above mentioned text is shown in table 5.

Table 5. Analysis of Machine Learning based approaches

Parameter / **Paper**	**Fattah et.al**[15]	**Yang et.al** [16]	**Desai et.al** [18]
Document Type	Single-document	Multi-document	Single-document
Language	Arabic, English	English	Hindi
Dataset	100 Arabic and 100 English documents	DUC-2004, DUC-2007	Hindi news articles
Machine Learning approach	Supervised learning	Un-supervised learning	Supervised learning
Technique	GA, FFNN, GMM, MR, PNN	Ranking based sentence clustering	SVM
Evaluation Parameter	Extrinsic	Both Intrinsic and Extrinsic	Extrinsic

IV. CONCLUSION AND FUTURE SCOPE

Automatic text extraction is a useful tool for information retrieval on the web and also it saves a lot of time of the user wandering for important information. This paper discusses the various approaches which the researchers are using to optimize the summary to the best possible. Also different languages and different document type are also considered for summary generation.

The generation of perfect summary is still not achievable. But with variations in the current approaches and considering a wider feature set might help to improve the quality of the summary.

REFERENCES

[1] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958.

[2] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques :," *Artif. Intell. Rev.*, vol. 47, no. 1, pp. 1–66, 2017.

[3] M. Hassel, "Evaluation of Automatic Text Summarization."

[4] "Extractive Summarization - Google Search." [Online]. Available: https://www.google.co.in/search?q=Extractive+Summarization&oq=Extractive+&aqs=chrome.1.69i57j69i59j69i61l3j0.7516j0j7&sourceid=chrome&ie=UTF-8. [Accessed: 25-Feb-2018].

[5] V. K. Gupta and T. J. Siddiqui, "Multi-document summarization using sentence clustering," in *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, 2012, pp. 1–5.

[6] Xiao-Ying Liu, Yi-Ming Zhou, and Ruo-Shi Zheng, "Measuring semantic similarity within sentences," in *2008 International Conference on Machine Learning and Cybernetics*, 2008, pp. 2558–2562.

[7] R. Ferreira *et al.*, "Assessing sentence scoring techniques for extractive text summarization," 2013.

[8] A. N. Gulati and S. D. Sawarkar, "A novel technique for multidocument Hindi text summarization," *2017 Int. Conf. Nascent Technol. Eng. ICNTE 2017 - Proc.*, 2017.

[9] L. Yang, X. Cai, S. Pan, H. Dai, and D. Mu, "Multi-document summarization based on sentence cluster using non-negative matrix factorization," *J. Intell. Fuzzy Syst.*, vol. 33, no. 3, pp. 1867–1879, 2017.

[10] G. Yang, D. Wen, Kinshuk, N.-S. Chen, and E. Sutinen, "A novel contextual topic model for multi-document summarization," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1340–1352, Feb. 2015.

[11] Z. Wu *et al.*, "A topic modeling based approach to novel document

automatic summarization," *Expert Syst. Appl.*, vol. 84, pp. 12–23, 2017.
[12] E. Baralis, L. Cagliero, N. Mahoto, and A. Fiori, "GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization," 2013.
[13] S. alZahir, Q. Fatima, and M. Cenek, "New graph-based text summarization method," in *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, 2015, pp. 396–401.
[14] C. Fang, D. Mu, Z. Deng, and Z. Wu, "Word-sentence co-ranking for automatic extractive text summarization," *Expert Syst. Appl.*, vol. 72, pp. 189–195, Apr. 2017.
[15] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 126–144, Jan. 2009.
[16] L. Yang, X. Cai, Y. Zhang, and P. Shi, "Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization," *Inf. Sci. (Ny).*, vol. 260, pp. 37–50, Mar. 2014.
[17] W. Li, W.-K. Ng, Y. Liu, and K.-L. Ong, "Enhancing the Effectiveness of Clustering with Spectra Analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 7, pp. 887–902, Jul. 2007.
[18] N. Desai and P. Shah, "AUTOMATIC TEXT SUMMARIZATION USING SUPERVISED MACHINE LEARNING TECHNIQUE FOR HINDI LANGAUGE."

Authors Profile

Ms. Sonam Gandotra pursed Bachelor of Computer Application from Govt. College for Women, Parade, University of Jammu, Jammu & Kashmir in 2011 and Master of Computer Application from University of Jammu, Jammu & Kashmir in year 2014. She is a gold medalist in MCA and has qualified NET-JRF and SET examinations. She is currently pursuing Ph.D. from Department of Computer Science & IT, Central University of Jammu, Jammu & Kashmir. Her area of interest are Text Mining and Natural Language Processing.

Dr. Bhavna Arora pursed Bachelor of Computer Science from Kurukshetra University and Post Graduated from Institute of Management Technology,Ghaziabad. She completed Ph.D from University of Jammu, in the year 2011.She has a total work experience of 21 years in industry and academia.She is presently working as Assistant Professor in Department of Computer Science & IT, Central University of Jammu,Jammu. She is a member of IEEE,ACM,CSI,SIE,ISTE,IETE. She has published more than 28 research papers in journals and conferences of national and international repute.She has attended international conferences sponsored by DST and has also received UGC-BSR grant for project.

