IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-6, Issue-4 E-ISSN: 2347-2693

Data Mining in IoT and its Challenges

Deepti Sehrawat[1*] and Nasib Singh Gill[1]

[1]Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India
[2]Department of Computer Science & Applications, Maharshi Dayanand University, Rohtak, Haryana, India

[*]*Corresponding Author: dips.scorpio@gmail.com*

Available online at: www.ijcseonline.org

Received: 10/Mar/2018, Revised: 17/Mar/2018, Accepted: 29/Mar/2018, Published: 30/Apr/2018

***Abstract*—** Internet of Things (IoT) has provided enormous opportunities to make prevailing smart environment by influencing the increasing ubiquity of Radio Frequency Identification Devices (RFID), wireless network, and sensor devices. Recently, a large number of industrial IoT applications have embarked their presence. Rapid technological growth introduces tremendous information on the network. Big Data is an idea to assemble huge amount of data from IoT enabled devices like sensors, actuators in IoT smart environment to help monitor specific conditions, procedures, and system performance. In this new generation, it becomes more challenging to extract most relevant information quickly and efficiently. To solve this problem, a data mining technique widely known as automatic text summarization may also prove to be fruitful. Text summarization creates summarized information from a large text corpus. Various latest techniques used for text summarization viz. Classification, Particle Swarm Optimization, Genetic Algorithms, clustering, neural network and various hybridized approaches are presented in this paper. The latest and relevant algorithms may be customized in the context of IoT applications. This paper is aimed at reviewing these techniques and also discusses the challenges as well as other related research issues.

***Keywords*—** Data mining in IoT, challenges, Multilingual text summarization, clustering, particle swarm optimization

I. INTRODUCTION

A number of smart applications and smart environments are now evolving with the Internet of Things (IoT). Presently, it is not so easy to quickly and efficiently forecast or thwart urgent health conditions in real-time. This is because, physician analyse the obtained health data offline. Due to huge demand of health data in smart environment the sensors are likely to be congested [1]. Nowadays, the data available on the network has extended such a huge size that it has turned out to be unfeasible to efficiently extract valuable information manually. The enormous amount of data in IoT enabled smart environment is thought of high business value. In order to extract information from this big data associated with IoT, efficient and optimised data mining approaches are required [2].

Text summarization gives a solution to this problem that automatically creates a relevant summarized information of multiple text documents [3]. "Automatic Text Summarization" technology is maturing day by day by providing solutions for information overloaded problems. Finding most relevant information in form of summary from a large volume of text is really a challenging task. So, in this fast and technologically advanced world, it becomes necessary to provide document summarization automatically. Automatic document summarization gives compress form of big data as a summary so to provide crux of the whole data to the users.

"Multi Document Summarization", on the other hand, is another similar technology in which summarization of multiple documents (set of documents), all together is performed. A similar approach of extractive text summarisation is presented in [4], which uses sentence based extraction.

Further, it is assumed that summarization is subjected to a similarity measure and optimized function. In Automatic summarization, multiple tasks are carried out and which directly affects the ability to generate high-quality summaries. If automatic summarization does not depend on particular language then it is called "language independent" and can be used for multilingual summarization [5]. Text summarization is said to be successful if results preserve the overall meaning of the entire data.

Data summarization techniques are also effectively applied to digital forensics by using indexing algorithms for searching digital evidences at the physical level to discover particular text strings [6]. A number of text summarization techniques are available these days like classification, clustering, association rule, latent semantics, probabilistic latent semantics, latent Dirichlet allocation and various soft computing approaches for example Particle Swarm

© 2018, IJCSE All Rights Reserved

Optimization (PSO), Genetic Algorithms (GA), Artificial Neural Networks (ANN), and etc. This paper presents numerous existing text summarization techniques along with their advantages and limitations.

A new data mining approach is edge computing that works on wireless, battery-powered, low power, smart sensing devices that sit at the edge points of the IoT. Edge computing/ edge mining can minimize the number of packets transfer in order to reduce data and energy requirements. Also, edge computing reduces the privacy risks by embedding information requests at the sensing point, controlling incorrect use. Authors in [7] presented a data-driven edge mining technique, called G-SIP in which data is transformed to a sparse form resulting in, reduction of packet transmissions, energy, and storage requirements. Due to huge sensitive and private data on the internet, data mining techniques are facing serious challenges. A new branch of data mining, Privacy-Preserving Data Mining (PPDM), comes into existence. PPDM, by protecting private and sensitive data performs data mining without the permission of data owners or providers [8].

Big data in IoT offers novel prospective to deliver analytical services to the citizens. However, ever increasing data on the internet has still a gap in uniting the present state-of-the-art in an integrated framework that would assist in the reduction of costs and facilitating new service types [9]. A number of approaches and models are available that deal with this huge data involved in smart IoT applications.

Organization of rest of the paper is as follows: Section 2 covers classification of text summarization techniques. In Section 3, various latest text summarization techniques and related work carried out over years by different researchers along with their main technologies and advantages. Section 4 focus on data mining models specific to IoT based smart applications. Various challenges to deal with huge IoT data are presented in section 5. Section 6 concludes the paper.

II. CLASSIFICATION OF TEXT SUMMARIZATION TECHNIQUES

Text summarization techniques can be characterized into 2 types, the first one is extractive summarization (Extraction) and the second is abstractive summarization (Abstraction), these are:

A. Extractive Summarization

In extractive summarization techniques, key sentences and paragraphs are carefully chosen from the documents and then data summarization is performed by concatenating these extracted parts. Statistical and linguistic methods are used to categorize important sentences. Authors in [5] introduced a multilingual extractive text summarization method by using a genetic algorithm. The proposed approach works on linear optimization of a number of sentence ranking measures. This proposed approach is named as MUSE. The main challenge of extraction techniques in text summarization is to select key representative sentences and some of the proposed selection techniques include “feature-base”, “cluster-base”, and “graph-base” methods. An improved hybrid “feature-base” and “cluster-based” approach that gives competent single-document summary is proposed by the authors in [10], by using Jaccard similarity measure (Jaccard coefficient). This approach makes use of simple real-to-integer values modulator to solve the discrete optimization problems and differential evolution algorithm for training and testing. Using fuzzy logic for extractive summarization has better results [11]. For the optimization of criteria functions, a discrete differential evolution algorithm is proposed and for sentence clustering, new criteria functions are also introduced [12].

B. Abstractive Summarization

On the other hand**,** abstractive summarization techniques use linguistic methods to interpret the meaning of original text and then describes it by generating shorter important information [13]. This approach comprises novel sentences, unseen in the original sources [12].One such approach is presented by the authors in [14], which makes a summary of an input text. The name given to this approach is conditional "Recurrent Neural Network" (RNN). A novel convolutional attention-based encoder is used for the conditioning which makes sure that at each and every step, the decoder focus on the suitable input words. The proposed model relies only on the learned parameters and can be trained with ease on large datasets. Several techniques are used for text summarization like, clustering, Particle Swarm Optimization (PSO), Genetic Algorithm [15], hybridized approaches and, etc. The automatic text summarization must generate a good summary with minimal redundant data in a short period of time [13].

III. TEXT SUMMARIZATION TECHNIQUES

Numerous text summarization techniques are available these days like classification, clustering, association rule, latent semantics, probabilistic latent semantics, latent dirichlet allocation and various soft computing approaches such as PSO, ANN, GA and etc.

A. Particle Swarm Optimization (PSO)

PSO is a latest heuristic search approach which is inspired from the swarming behavior of living inhabitants [15]. It is a population-based searching which moves from population to population in a single iteration with probable enhancement. This approach makes use of combined deterministic and probabilistic rules. Authors in [15] presented a PSO technique by applying statistical analysis. This proficient approach makes use of formal hypothesis testing and is as efficient as GA with significantly better computational efficiency. A similar study in [16] reveals that conventional algorithms

© 2018, IJCSE All Rights Reserved

have some limitations like slow convergence and sensitivity in initializing values. On the other hand PSO gives better results in complex and multi-peak problems and has outstanding performance in high-dimensional data clustering and also proves to be efficient in linearly non-separable datasets. PSO does not need any prior knowledge about number of clusters to be formed. Enhanced PSO by hybridizing other algorithms like K-means, KFC, Genetic Algorithms, Ant Colony Optimization, dimensionality reduction and feature selection can be effectively applied for clustering high-dimensional data [16].

B. Clustering

One most popular data mining technique is data clustering where an unlabeled dataset is partitioned into groups called clusters having similar objects with respect to some similarity measure. A sentence clustering algorithm working on a graph model by utilizing statistic similarities and linguistic treatment is proposed [3]. It is a "multi-document summarization" technique, to extract important information from documents by avoiding information redundancy and diversity. A similar improved extractive summarization approach is proposed by the authors in [4] by applying a continuous sentence clustering. It is a two-step process in which, first sentence clustering is performed, and then in the second step on individual cluster, representative sentences are defined. A discrete differential evolution algorithm that optimizes the objective function is proposed. It is proved, that similarity measures have great impact on the results obtained in summarization. A similar study in [17] proposed a non-specific summarization approach for multi-documents. Proposed approach relies on sentence clustering by introducing 5 methods of clustering to optimize several characteristics of "intra cluster similarity", "inter cluster dissimilarity", and their combinations. A modified discrete PSO is proposed for clustering problem.

C. Genetic Algorithm (GA)

A genetic algorithm is a "population-based" technique and is intuitive, easily implementable, and able to efficiently solve complex (extremely non-linear mixed integer optimization) problems. The main limitation of GA is its computational cost, it is very expensive as compared to others [15]. Authors in [5] introduced a multilingual extractive text summarization method by using a genetic algorithm which relies on linear optimization of numerous measures to rank sentences. This proposed approach is named as MUSE.

D. Neural Network

Neural network needs training in order to learn some specific characteristics. In text summarization, neural networks are trained to summarize text corpus through learning sentence characteristics. Modification of neural network is performed so as to simplify and combine the substantial characteristics which are obvious in summary. Lastly, the modified neural network filter the text corpus to summarize news articles. The proposed approach makes use of training by individual readers according to their style and selected features can be reformed to as per reader's necessities and choice [18].

E. Suffix Stripping

Removing suffixes/ suffix stripping is proposed as an implementation in BCPL. It efficiently works by automatically stripping texts and consider complex suffixes as compounds which are made up of simple suffixes. It progresses by eliminating simple suffixes in several steps depending upon the type of the leftover stem, which typically includes a measure such as its syllable length [19].

F. Integer Linear Programming Problem

Generating summarized data using unsupervised sentence extraction approach is proposed in [20]. It can be used for both, single document summarization and multiple document summarization. The proposed approach modeled text summarization as an integer linear programming problem. Key sentences are retrieved directly by covering main contents of the original documents. Three properties are optimized in this model on which it focusses, these are extraction of relevant information, irredundant information retrieval and bounded length summary.

A number of text summarization techniques are available these days. Table 1 enlists major techniques used in recent works along with their advantages and limitations. These limitations can be taken as a future research area.

Table 1. Important Publications and techniques used in their approach

Pub. Year	Title of Paper	Authors	Techniques Used	Advantages
2017	"Topic and user based refinement for competitive perspective identification" [21]	J. Lin, W. Mao and D. Zeng	Supervised Refined Perception Classifiers with document-topic distributions mined from texts. User-based bootstrapping method, semi-supervised in nature.	It helps to select good quality classified texts from an unlabeled online corpus.
2017	"Accelerating Topic Exploration of Multi-Dimensional Documents" [22]	H. Wen-Jing, L. You and L. Z. Qi	Indexes are created then for documents subsets topic models are pre-computed.	Gives better and improved solution to identify correct sets of documents. Can be parallelized and pre-computing is possible.

© 2018, IJCSE All Rights Reserved

2016	"Abstractive Sentence Summarization with Attentive Recurrent Neural Networks" [12]	S. Chopra, M. Auli and A. M. Rush	Conditioning using encoder	Simple and effective automatic summaries can be generated.
2016	"Topic discovery and future trend forecasting for texts" [23]	J. L. Hurtado, A. Agarwal, and X. Zhu	Association rule mining and Ensemble topic forecasting	Has better performance
2016	"Knowledge Based Summarization and Document Generation using Bayesian Network" [24]	S. Malviya and U. S. Tiwary	Bayesian Network	50% of relevance of the summary with the source.
2016	"Forecasting Emerging Trends from Scientific Literature based on Keyword Extraction and Prediction" [25]	K. Asooja, G. Bordea, G. Vulcu and P. Buitelaar	Regression models, Spearman Rank Correlation, Average Root Mean Square	Future Forecasting is done.
2015	"Summarizing large text collection using topic modeling and clustering based on MapReduce framework" [26]	N. K. Nagwani	K-Means clustering and Latent Dirichlet Allocation	Faster and powerful for analysing Big Text Data.
2014	"Text Summarization Using Neural Networks " [16]	K. Kaikhah	Neural network	Customized based as per requirements.
2014	"A multi-document summarization system based on statistics and linguistic treatment"[3]	R. Ferreira, L. D. S. Cabral, F. Freitas, R. D. Lins, G. D. F. Silva, S. J. Simske and L. Favaro	Sentence Clustering which is based on graph model using statistic similarities and linguistic treatment.	Briefly extracts key parts from large text corpus from multiple documents by avoiding information redundancy and diversity.
2013	"Integrating document clustering and topic modeling" [27]	P. Xie and E. P. Xing	Clustering	Able to differentiate topics into group specific ones and group-independent ones.
2013	"Evaluation of text document clustering approach based on particle swarm optimization" [28]	S. Karol and V. Mangat	Hybrid of K-Means and PSO algorithm (KPSO) and Hybrid of FCM and PSO algorithm (FCPSO)	Better performance
2012	"Temporal Corpus Summarization Using Submodular Word Coverage"[29]	R. Sipos, A. Swaminathan, P. Shivaswamy and T. Joachims	K-neighbor, Greedy submodular	Offer monotone sub-modular goals for generating useful summaries of temporal documents over time.
2011	"A frequent keyword-set based algorithm for topic modeling and clustering of research papers" [30]	K. Shubhanker, A. P. Singh and V. Pudi	Hierarchical clustering, Modified Page rank Algorithm	Fast, effective and scalable
2011	"A Time-Dependent Topic Model for Multiple Text Streams"[31]	L. Hong, B. Dom, S. Gurumurthy and K. Tsioutsiouliklis	Latent Dirichlet Allocation, Gibbs sampling	Can be applied to unseen multiple temporal data to get coherent topics.
2010	"A new Approach to Improving Multilingual Summarization using a Genetic Algorithm"[5]	M. Litvak, M. Last and M. Friedman	Genetic Algorithm and Sentence classification measures.	A single model can be used for more than one language.
2008	"Automatic Text Summarization" [32]	M. A. Fattah and F. Ren	Genetic algorithm (GA) and mathematical regression (MR)	Used feature extraction criteria for several features based on the used language and the text type.
2006	"Probabilistic topic decomposition of an eighteenth-century American newspaper" [33]	D. J. Newman and S. Block	Probabilistic Latent Semantic Approaches (pLSA)	Able to find arbitrary levels detail, and topic hierarchy for historian's specific interests.
2005	"Discovering Evolutionary theme patterns from text: An exploration of temporal text mining" [34]	Q. Ma and C. X. Zha	Clustering, SVM Classification and Kullback- Leibler Divergence probabilistic approach	Handle all types of text stream data including a temporal stream.
2003	"Mining Topic-Specific concepts and definitions on the web" [35]	B. Liu, C. W. Chin and H. T. Ng	Agglomerative clustering and PERL Language	Help web users to learn an unfamiliar topic systematically and in-depth.

IV. DATA MINING IN IOT

Big Data collects, process and analyse huge relevant data from internet enabled devices like sensors attached to monitor specific conditions, usage and system performance. Now a days, a number of data mining applications come into existence. This section covers data mining approaches in IoT.

By deploying IoT all the way through the water supply set-up along with usage details of different consumers, a schematic IoT application for big data collection is proposed. The proposed approach uses "Wireless Sensor Network" technologies for downstream and upstream data collection. Downstream data intend to deliver information about water usage details along with its performance to the clients' on the other hand upstream data is for "Automated Meter Reading" systems. Finally, converging to big-data collection and here data mining works by identifying: "local and system performances including pressure and flow", "non-revenue and illegitimate water consumption" and "locations and quantity of water breaks and water losses". This new approach allows both consumers along with utilities to proactively manage their water usage [36]. A similar approach for smart agriculture proposed a multidisciplinary

© 2018, IJCSE All Rights Reserved

model called, AgroCloud. In this model farmers, vendors and agencies are required to be registered to AgroCloud module. For this, it provides a MobileApp module. It stores huge agriculture data all about farmers, vendors, agencies, government schemes, environment updates and properties of periodic soil. Through IoT, sensors are attached so as to monitor the properties and conditions of soil and environment, which are sent to AgroCloud. To find best suitable fertilizers requirements, water requirements, total production, market requirements and etc. big data analysis is done [37].

A six-layer architecture for data mining in IoT applications works for devices, raw data, service, security, privacy, and standards [2].

A Parallel programming model by designing a dynamic multisource data mining model is proposed by the authors in [38] which deal with dynamic and stream data. A similar approach in [39] gives parallel association rule mining approach by using a k-means algorithm.

To find correlation among multi-faced data, a new approach can simultaneously deal with this data by transforming multi-faced data which is of high dimension into a data of low dimension. The mining is done by finding the structure/ correlation within the data that exists [40].

Constrained IoT environment has laid down various challenges for the successful implementation of big data mining concepts.

V. CHALLENGES IN IOT AND BIG DATA ERA

In big-data era, rapid growth of smart applications posed numerous challenges. After deep analysis of various existing text summarization techniques, it is found that the most important challenge in Internet of Things and big-data era is extracting useful information from huge text corpus by protecting data and providing privacy and security to the secret and private data.

Key challenges in IoT are as follows:

1) Fast and effective data mining approaches for huge data corpus to read and write.
2) Integration of heterogeneous, noisy, uncertain, incomplete, fuzzy, structured and semi-structured and unstructured data sources and data types.
3) Communication via diverse device types and systems and furthermore need to mine webpage data.
4) Need to extract complex knowledge after analyzing properties of the data and finding association among them.
5) Need to develop efficient and well-defined data mining framework which considers data security, data privacy, big data, and data sharing of prime importance.
6) Need of enhanced parallel programming model by designing a dynamic multisource data mining model.

Most of the present data mining algorithms are not adapted to parallel platform and requires attention of researchers in this area.

VI. CONCLUSION and Future Scope

To assess, process and extract important valuable information from huge corpus in IoT enabled smart environment a new concept, edge mining, is used. This new emerging area of IoT has not been realized so far since technological collaborations and smart environment is still developing and is in infancy stage. Big data mining of heterogeneous networks is a growing area. However, there are some challenges associated with big data mining like varied, unrelated, overlapping information and summarizing multi-documents. Consequently, it becomes more important to shrink the redundant information present in the data and which also reflects in the summaries too. Thereafter, extracting common sentences from it and refining final results. In recent years, numerous supervised and unsupervised approaches have been proposed which are based on syntactic or statistical features. The need for quick and efficient data mining in IoT dealing with fuzzy heterogeneous data and communicating over different types of devices is a main challenge in IoT and big data.

REFERENCES

[1] J. J. Kang, S. Adibi, H. Larkin, and T Luan, "Predictive data mining for converged internet of things: A mobile health perspective", In *Telecommunication Networks and Applications Conference (ITNAC), 2015 International*, pp. 5-10. IEEE, 2015.

[2] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges", *International Journal of Distributed Sensor Networks, Vol.* 11, No. 8,2015.

[3] R. Ferreira, L. D. S. Cabral, F. Freitas, R. D. Lins, G. D. F. Silva, S. J. Simske, and L. Favaro, "A multi-document summarization system based on statistics and linguistic treatment", Expert Systems with Applications, Vol. 41, No. 13, pp: 5780-5787, 2014.

[4] R. M. Aliguliyev, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", *Expert Systems with Applications* 36, no. 4, pp.: 7764-7772, 2009.

[5] M. Litvak, M. Last and M. Friedman, "A new approach to improving multilingual summarization using a genetic algorithm", In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pp. 927-936. Association for Computational Linguistics, 2010.

© 2018, IJCSE All Rights Reserved

[6] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results", *Digital investigation* 4, pp. 49-54, 2007.
[7] E. I. Gaura, J. Brusey, M. Allen, R. Wilkins, D. Goldsmith and R. Rednic, "Edge mining the internet of things", *IEEE Sensors Journal* Vol.13, No. 10, pp.: 3816-3825, 2013.
[8] X. Li, Z. Yan and P. Zhang, "A review on privacy-preserving data mining", In proceedings of *2014 IEEE International Conference on Computer and Information Technology (CIT)*, pp. 769-774, 2014.
[9] F. Xhafa, L. Barolli, A. Barolli, and P. Papajorgji., "Modeling and processing for next-generation big-data technologies", *Cham: Springer International Publishing* (2015).
[10] A. Abuobieda, N. Salim, Y. J. Kumar and A. H. Osman, "An improved evolutionary algorithm for extractive text summarization", In *Asian Conference on Intelligent Information and Database Systems*, pp. 78-89. Springer, Berlin, Heidelberg, 2013.
[11] P. D. Patil and N. J. Kulkarni, "Text summarization using fuzzy logic", *International Journal of Innovative Research in Advanced Engineering (IJIRAE), Vol.* 1, 2014.
[12] R. Alguliev and R. Aliguliyev, "Evolutionary algorithm for extractive text summarization", *Intelligent Information Management*, Vol. 1, No. 02, pp. 128-138, 2009.
[13] V. Gupta and G. S. Lehal, "A survey of text summarization extractive techniques", Journal of emerging technologies in web intelligence, Vol. 2, No. 3, pp.: 258-268, 2010.
[14] S. Chopra, M. Auli and A. M. Rush, "Abstractive sentence summarization with attentive recurrent neural networks", In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.: 93-98, 2016.
[15] R. Hassan, B. Cohanim, O. D. Weck and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm", In *46th AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference*, pp. 1897. 2005.
[16] A. A. A. Esmin, R. A. Coelho and S. Matwin, "A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data", *Artificial Intelligence Review* Vol.44, No. 1, pp.: 23-45, 2015.
[17] R. M. Aliguliyev, "Clustering Techniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization", *Computational Intelligence* Vol. 26, No. 4, pp.: 420-448, 2010.
[18] K. Kaikhah, "Text summarization using neural networks" Faculty Publications, Texas State University, 2004.
[19] M. F. Porter, "An algorithm for suffix stripping", *Program* Vol. 14, No. 3, pp.:130-137, 1980.
[20] R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova and C. A. Mehdiyev, "MCMR: Maximum coverage and minimum redundant text summarization model", *Expert Systems with Applications* Vol. 38, No. 12, pp.: 14514-14522, 2011.
[21] J. Lin, W. Mao, and D. Zeng, "Topic and user based refinement for competitive perspective identification." *Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on*. IEEE, 2017.
[22] H. W. Jing, L. You, and L. Z. Qi, "Accelerating Topic Exploration of Multi-Dimensional Documents", In Parallel and Distributed Processing Symposium Workshops (IPDPSW), 2017 IEEE International, pp. 1520-1527. IEEE, 2017.
[23] J. L. Hurtado, A. Agarwal, and X. Zhu, "Topic discovery and future trend forecasting for texts", Journal of Big Data, Vol. 3, no. 1, 2016.
[24] S. Malviya and U. S. Tiwary, "Knowledge Based Summarization and Document Generation using Bayesian Network", Procedia Computer Science 89, pp. 333-340, 2016.
[25] K. Asooja, G. Bordea, G. Vulcu, and P. Buitelaar, "Forecasting Emerging Trends from Scientific Literature", In LREC. 2016.
[26] N. K. Nagwani, "Summarizing large text collection using topic modeling and clustering based on MapReduce framework", Journal of Big Data Vol. 2, no. 1, 2015.
[27] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling", In the proceeding of the Twenty-Ninth conference on Uncertainty in Artificial Intelligence (UAI), pp. 694-703, 2013.
[28] S. Karol and V. Mangat, "Evaluation of text document clustering approach based on particle swarm optimization", Open Computer Science 3, no. 2 (2013): 69-90.
[29] R. Sipos, A. Swaminathan, P. Shivaswamy, and T. Joachims, "Temporal corpus summarization using submodular word coverage", In Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 754-763. ACM, 2012.
[30] S. Kumar, A. P. Singh and V. Pudi, "A frequent keyword-set based algorithm for topic modeling and clustering of research papers", In Data Mining and Optimization (DMO), 2011 3rd Conference on, pp. 96-102. IEEE, 2011.
[31] L. Hong, B. Dom, S. Gurumurthy and K. Tsioutsiouliklis, "A time-dependent topic model for multiple text streams", In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 832-840. ACM, 2011.
[32] M. A. Fattah and F. Ren, "Automatic text summarization", *World Academy of Science, Engineering and Technology*, Vol. 2, No. 1, 2008.
[33] D. J. Newman and S. Block, "Probabilistic topic decomposition of an eighteenth-century American newspaper", *Journal of the Association for Information Science and Technology* Vol. 57, No. 6, pp. 753-767, 2006.
[34] Q. Mei and C. X. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining", In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 198-207. ACM, 2005.
[35] B. Liu, C. W. Chin, and H. T. Ng, "Mining topic-specific concepts and definitions on the web", In *Proceedings of the 12th international conference on World Wide Web*, pp. 251-260. ACM, 2003.
[36] D. Koo, K. Piratla and C. J. Matthews, "Towards sustainable water supply: Schematic development of big data collection using internet of things (iot)", *Procedia engineering*118 (2015): 489-497.
[37] H. Channe, S. Kothari and D. Kadam, "Multidisciplinary model for smart agriculture using internet-of-things (IoT), sensors, cloud-computing, mobile-computing & big-data analysis", *Int. J. Computer Technology & Applications*, Vol. 6, No. 3, pp.:374-382, 2015.

© 2018, IJCSE All Rights Reserved

[38] X. Wu, X. Zhu, G. Q. Wu and W. Ding, "Data mining with big data", *IEEE transactions on knowledge and data engineering*, Vol. 26, No. 1, pp.:97-107, 2014.
[39] X. Wu and S. Zhang, "Synthesizing high-frequency rules from different data sources", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 2, pp.:353-367, 2003.
[40] C. Dong, Q. Xiuquan, J. Gelernter, L. Xiaofeng and M. Luoming, "Mining data correlation from multi-faceted sensor data in the Internet of Things", *China Communications*, Vol. 8, No. 1, pp.:132-138, 2011.

Authors Profile

Ms. Deepti Sehrawat has passed Master of Science and Applications from Department of Computer Science & Applications, M. D. University, Rohtak, India in 2008 and Master of Philosophy from C. D. L. University, Sirsa in 2009. She is currently pursuing Ph. D under the supervision of renowned academician and researcher – Professor Nasib Singh Gill of M. D. University. She has published more than 20 research papers in reputed National and International Journals and Conference Proceedings including IEEE. Her main research work focuses on IoT, Lightweight Cryptography Algorithms, Network Security and Privacy, Big Data Analytics and Data Mining. She has 8 years of teaching experience.

Dr. Nasib Singh Gill is at present senior most Professor of Department of Computer Science & Applications, M. D. University, Rohtak, India and is working in the Department since 1990. He has earned his Doctorate in Computer Science in the year 1996 and carried out his Post-Doctoral research at Brunel University, West London during 2001-2002. He is a recipient of Commonwealth Fellowship Award of British Government for the Year 2001. Besides, he also has earned his MBA degree. He has published more than 245 research papers in reputed National & International Journals, Conference Proceedings, Bulletins, Edited Books, and Newspapers. He has authored seven books. He is a Senior Member of IACSIT as well as a fellow of several professional bodies including IETE and CSI. He has been serving as Editorial Board Member, Guest Editor, Reviewer of International/National Journals and a Member of Technical Committee of several International/National Conferences. He has guided so far 8 Ph.D. scholars as well as guiding about 7 more scholars presently in the areas – IoT, Information and Network Security, Computer Networks, Measurement of Component-based Systems, Complexity of Software Systems, Decision Trees, Component-based Testing, Data mining & Data warehousing, and NLP.

© 2018, IJCSE All Rights Reserved