Information Extraction from Multifaceted Unstructured Big Data

Kiran Adnan, Rehan Akbar, Khor Siak Wang

Abstract*: In the era of digital globalization, huge volume and variety of data are being produced at a very high rate. Every day, the world is producing around 2.5 quintillion bytes of data. According to IDC, by 2020, over 40 zettabytes of data will be generated and reproduced. Digital data have become a deluge, overwhelming in every field of information technology (IT), business, science and engineering. These fields are shifting to smart and advanced technologies such as smart manufacturing industries, data-aware medical sciences, and other smart applications. These applications are facilitating the industries in context of data-driven decision making, big data storage, and complex analysis of large data sets. Also, these applications are contributing to generate big data deluge where a variety of data necessitate the industries to use advanced IT approaches. 95% of the digital universe is unstructured data. It is rich data as it contains information that can play a vital role to improve big data analytics. The heterogeneity, complexity, lack of structured information, poor quality and scalability of unstructured data generates difficulties in adapting traditional information extraction techniques. Information extraction can play a vital role in transformation of unstructured data into useful information. A multistep pipeline with data preprocessing steps, extraction methods and representation are utmost requirement to improve the unstructured data analytics. In this regard, this paper presents a short review of information extraction process w.r.t. input data type, extraction methods with their corresponding techniques, and representation of extracted information. The issues with unstructured data and the challenges to information extraction from multifaceted unstructured big data as well as the future research directions have also been discussed

***Index Terms*: unstructured data, information extraction, big data, unstructured data analysis**

I. INTRODUCTION

Digitalized and interconnected world is generating huge volume and variety (structured, semi-structured and unstructured) of big data on daily basis. Unstructured Data is the most important because it comes without any structure or formal data model which makes it difficult to process, manage and store. Unstructured data growth rate is much higher than structured data. 95% of digital universe will be unstructured in 2020 and it is getting double in every two years [1]. Information Extraction (IE) from a variety of data types is very challenging due to the issues of unstructured data. Several IE techniques to extract different type of information from different types of data i.e. text, image, audio and video has been discussed here. Named Entity Recognition, Coreference resolution, relation extraction and event extraction are the sub-tasks of IE from text. Several rule-based, knowledge-based, and learning-based techniques are discussed to extract information from text data. Different type of information can be extracted from images as well such as feature extraction which includes color, shape, texture and edge detection, object and character recognition, text extraction and semantic information extraction. IE from audio data includes speech recognition, subtitling, sematic annotation, face recognition and many more. Although, only some prominent techniques to extract information from audio data are discussed here. Video summarization is a technique to extract a summary of visual content in static and dynamic manner. All the above-mentioned solutions are limited to specific domains, data types, languages and other features.

This paper reviews IE process from three aspects as described in fig 1 below. Input aspect describes the data type for IE process. There are four input data types has been discussed here i.e. text, image, audio, and video. From each data type input, Extraction method has been elaborated along with their techniques. And what information these techniques extract, has also been discussed. For example, in text input, Entity

Extraction is an extraction method and supervised, unsupervised or hybrid approaches are the techniques to extract entities from text. At the end, entities are represented as output.

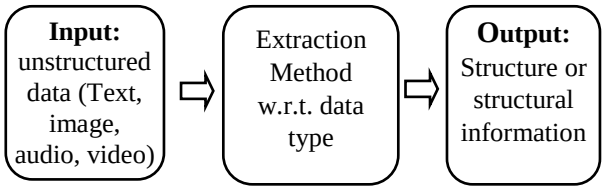

Fig 1: Information Extraction Aspect

The process of automatic IE from unstructured data opens new ways to make its processing and management more strong. In this regard, this paper identifies the issues of IE process that comes with unstructured data in order to optimize the performance issues. This review highlights the need of multistep IE pipeline to handle the unstructured data deluge. The remaining paper is organized as follows. Section 2 briefly reviews the IE techniques and methods to extract different type of information from the variety of unstructured data. In section 3, the issues of unstructured data and challenges to IE have been discussed. Conclusion is

Revised Manuscript Received on August 19, 2019.

Kiran Adnan, Faculty of Information & Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia.
Rehan Akbar, Faculty of Information & Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia.
Khor Siak Wang, Faculty of Information & Communication Technology, Universiti Tunku Abdul Rahman, Kampar, Malaysia.

presented in section 4.

II. Information Extraction (IE) & Unstructured Data

People and machines are producing data at very high rate than ever before. The volume and variety of data being produced brings more challenges in identifying useful information from it. As unstructured data are 90% of the huge deluge of data, it becomes more difficult to identify and extract the required information from it at the right time. Information extraction is a process to extract the content and context from unstructured data [2]. While the unstructured data growth rate is far high, a lot of challenges needs to be considered to improve the efficiency and usefulness of unstructured data. Scalability, complexity and heterogeneity of unstructured data appears as main challenges to harvest useful information. Organization of unstructured and heterogeneous big data is a technical challenge in terms of information extraction and presentation. Transformation of unstructured data into structured format for better representation are the big questions [3]. Efficient and accurate automatic transformation of unstructured data is utmost need to improve the analytical process. The following subsections review IE techniques for different data types i.e. text, image, audio and video respectively. For each data type, extraction methods along with the techniques have been discussed.

A. Information Extraction from Text data

The core idea behind IE from text is to transform unstructured data into structured by annotating semantic information. IE systems can be summarized by several dimensions such as level of logical structures, nature of data source, linguistic components and domain criteria. Some prominent research work has been discussed here about IE from text without focusing any single domain. Low-level structure identification in the process of IE tasks is categorized into Named Entity Recognition NER (descriptive information about named entities), Relation extraction RE (identifying relation between entities), and Event extraction EE (identifying events in text) [4]. Natural Language Processing, Machine Learning, Text mining and Computational Linguistics are helping IE process to bridge the semantic gap, and in extracting and representing the most relevant information. However, a huge volume of multifaceted unstructured data makes IE process more challenging.

Extraction methods use rule-based, machine learning and hybrid approaches to extract information from text. Rule-based methods for NER uses lexico-syntactic patterns and semantic constraints to identify the occurrence of similar entities whereas learning based methods uses machine learning to extract named entities and its classification. Learning based methods can be supervised such as Hidden Markov Model (HMM), Maximum Entropy Model (MaxEnt), Support Vector Machine (SVM) and Conditional Random Fields (CRF), unsupervised such as clustering (hard and soft) and semi-supervised such as bootstrapping. Supervised and unsupervised approaches used a large amount of training data to achieve high performance but semi-supervised uses both labeled and unlabeled corpus with a small degree of supervision [5]. Entity extraction from unstructured text documents and annotating by the concept identified, improves the effectiveness of results. However, it was considered a resource intensive procedure to create annotated corpus. Various techniques are introduced to automatically label the corpus for training the model. But unfortunately, Automatic Labeling of training data includes noise in terms of missing or wrong labels that causes incompleteness and inconsistency [6]. In this regard, Machine Learning and Deep Learning solutions are more reliable to make this task labor and cost effective. The need is to identify most effective techniques for the problem at hand. Hybrid neural network models, joint models have been proposed to remove the sequential information extraction process for NER and RE. These models are capable to handle the error propagation by dividing the NER and RE task as two different simultaneous tasks. Joint models using deep learning are more effective for feature extraction from unstructured documents [7]. The Heterogeneity and complexity of unstructured data makes this task more daunting. Traditional methods to identify entities are not sufficient due to the multidimensionality, heterogeneity, and complexity of unstructured big data.

Natural Language Processing (NLP) tools and techniques have a significant role in the domain of IE. The advancements in NLP systems was started from word segmentation, part-of-speech (POS) tagging, and morphological analysis to shallow parsing, NER, and relation extraction among entities and terms. Research on relation extraction are categorized into identifying universal schema and collaborative filtering. Generally, Relation Extraction aims to discover structured and semantic relationship among identified named entities. In this regard, several supervised with feature-based and kernel-based approaches, unsupervised and semi-supervised methods have been proposed. Supervised methods uses labeled datasets for training the models to identify the patterns of relation types [8], [9]. Extracting different features like syntactic, dependency, and semantic, are addressed using linguistic pattern learning [10]. But these methods need annotated corpora to train the model to identify pair of entities. To overcome this limitation, unsupervised and semi-supervised methods were introduced that uses heuristic rules or different type of clustering algorithms to identify relations between entities from large unlabeled corpus [11], [12]. Several hybrid approaches have also been introduced to maximize the efficiency for different languages, but linguistic components and domain criteria affects the efficiency and accuracy of these approaches. NER and RE can be improved by adopting linguistic analysis with NLP [13].

There are different in-practice techniques for EE; Data-driven (focus on specific features such as words, n-grams, weights etc.), Knowledge-driven (Lexico-syntactic patterns and lexico-semantic patterns) and Hybrid approaches. Hybrid approaches are compromising techniques between these two approaches to minimize the effort and to improve the performance. Although, high expertise are

required to develop a hybrid approach [14]. In this regard, topic identification from unstructured documents using hybrid approach was introduced for Chinese language that used segmentation rules and statistical methods to identify semantic relations among terms. The results in the form of knowledge graph showed pretty good results [15].

IE from human language text is different for each language. But IE is easier for rich morphological languages like Russian and English. Several solutions for various languages such as Urdu [16], Malayalam [17], and Kurdish [18] are proposed in literature.

IE from text highly depends on the domain and language. Languages makes IE task more challenging and removing the linguistic barrier can improve the efficiency of IE process. In this regard, Machine translation is a solution which translate text from one language to another language. Once the text is translated, auto coding and indexing is applied to extract terms from text [19]. Lexical and semantic features, syntax and logical relations are categorized into language independent and language dependent features respectively, can be combined to extract domain-related information from digitized textbooks of different languages [20]. Data extraction, syntactic and semantic analysis, classification, inference rules and representation into XML are the steps to add structure to unstructured text [21]. But XML is not capable to represent complex unstructured data in fully structured manner. Structured data is more efficient due to the process of normalization in RDBMS. In this regard, a general process to extract information from unstructured text and stores it in database consists of five tasks i.e. segmentation, classification, association, normalization and deduplication [22]. The proposed IE process had limitation of a single data type and accuracy of these methods varies depending on the complexity and quality of data. Different domains' data, Languages and heterogeneous data sources make this task more complicated. Portability of IE process is challenging task without understanding the flow of IE process irrespective of domain and language dependency.

B. Information Extraction from Image data

Digital media and globalization giving rise to the visual unstructured data which is rich in content and context. However, extracting linguistic descriptions, semantic and visual features is a challenging paradigm in terms of improved efficiency and accuracy. Feature extraction, text extraction, character recognition, scene understanding, and geospatial IE are some prominent areas. Feature extraction is helpful in identifying the visual objects in images. Different classification and segmentation approaches have been proposed to extract useful objects from images. Scale Invariant Feature Transform (SIFT) is used for feature extraction in images. Target detection using segmentation and SIFT has shown average classification accuracy up to 90.99% [23]. Color (histogram, color coherence vector and color correlogram), shape (contour and region based methods) and texture (spatial and spectral extraction methods) features are most frequently used features for IE [24].

Text and characters inside the images contain a vast array of information. Text extraction from images facilitates to extract useful information but also facing several challenges such as detection and recognition of text due to language, size, font, orientation, contrast, complex colored and textured background. Several approaches have been proposed to extract this information from images, but all the approaches are either domain dependent or language dependent. There is no single unified technique to extract textual information from images for all applications [25]. The advantages of IE from images are efficiency, less complexity and less time consuming but when image is noisy, one cannot take advantage without noise removal prior to IE [26]. Top-down, bottom-up and combined approaches are hierarchical strategies to understand the scene contextually and semantically. In extracting meaningful features, many challenges lies in dynamic scene understanding such as type and position of images, scene motion, illumination changes, static and dynamic occlusions, type speed and pose of objects, camera synchronization and handover, event complexity and handling dynamic scenes [27].

C. Information Extraction from Audio data

Companies like call centers and music are the major sources which generate a huge volume of audio data. Audio data has different hidden information as compared to text and images. Different type of information can be extracted from these data to help predictive and descriptive analytics in big data. Automatic speech recognition (ASR) is mostly used for speech to text conversion. Several feature extraction techniques in speech recognition such as Linear Predictive Analysis (LPC), Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Predictive Coefficients (PLP), Mel-frequency Cepstral Coefficients (MFCC), Power Spectral Analysis (FFT), Relative Spectra filtering of log domain coefficients (RASTA) and First Order Derivative (DELTA) are used. The efficiency and effectiveness of these techniques depends on the languages used. However, hybrid feature extraction leads to more accurate results and better performance [28], [29]. The combination of SVM and Artificial Neural Network (ANN) techniques leads to reasonable performance for event detection. Some events such as drums, hammering and laughing are not handled by this combination [30]. Sound or acoustic event extraction convert audio signals into symbolic descriptions. Long short term memory recurrent neural network has been used to concentrate on mono channel audio by ignoring overlapping voices [31] but limited to less noisy environmental audio.

Several low-level and high-level features are extracted from audio data but the use of mid-level feature to extract rhythm-related information showed 93% accuracy as compared to baseline methods. The efficiency of mid-level feature extraction can also be helpful to different applications like automatic music sequencing, music database navigation, mash-up creation and complement systems [32]. MPEG-7 low-level features and interlaced derivate pattern (IDP) can be used to produce a set of features. In this regard, relevant feature extraction from unstructured data with nonlinear combination of sets of features and decision making based on three machine learning classification approaches was

proposed to remove uncertainty [33].

Extracting linguistic information through transcription correction is a letter to sound conversion task to improve the speech synthesis and recognition [34] but techniques applied only to Amharic, Hindi and Tamil languages. This field is facing a lot of challenges such as understanding speech in different languages with higher accuracy, overlapping nature of content, non-exclusive audio classification, feature selection algorithm.

Semantic IE to extract music score and text information using segmentation and classification is achieved by analyzing arbitrary soundtracks and timestamp the occurrence of music and speech [35]. An IE approach is required for integrated detection and verification from speech which can be useful for speech analysis, speech recognition, speaker and language recognition [36]. Although processing conversational speech, music, and monitoring conversations in call center are different examples in which IE is facing certain challenges such as noise in background, overlapping of words, focus on voice in crowd and language understanding.

D. Information Extraction from Video data

Face recognition in video i.e. face processing, construction of graphs and phonogram among characters, interaction score computing, character identification and visual graph construction scenario are the components to extract useful information from video [37]. Semantic concept detection provides semantic labels for videos at different levels. The combination of annotation and high-dimensional features is an active research area which uses various projections, data reductions and generalization techniques to avoid false positives and missed detections. Video summarization in terms of extracting speaker information and face recognition to identify the speakers in video, is a field where face recognition using Non-negative Matrix Factorization (NMF) technique has been used to decompose the facial area and after face recognition, GMM-SVM (Gaussian mixture mode based support vector machine) approach has been applied to identify the speaker [38]. This approach has given efficient results and worked better if video has target person's face but no voice and vice versa. Temporal IE, for less scene changing video to extract common features, can be used to assess the quality of video where several quality parameters considers for every frame to extract information about the variety of difference between two successive frames [39]. Diversity of unstructured content, lack of structure, and low quality in unstructured video content makes video summarization task more complex [40].

Video segmentation uses visual cues that are helpful to extract useful information to summarize sports videos automatically. Experimental study has been conducted for the summarization of event segmentation system on basketball broadcasting video [41]. Automatic subtitle generation is an important field of IE and facing several challenges such as linguistic barrier and accent issues. This process includes audio extraction, speech recognition using decoder and text synchronization to video [42]. Visual feature extraction from video, like shape, appearance of jaw and lip, is helpful to extract phonetic and visemic information using audio-visual speech recognizer. Audio-visual speech synthesis and recognition uses speech to text and text to speech conversion technique [43]. Video IE is facing challenges such as background and foreground extraction, speech to text and text to speech conversion, automatic labeling, text information available in video, learning semantics from features.

III. CHALLENGES TO INFORMATION EXTRACTION FROM UNSTRUCTURED DATA & RESULTS

Many enterprises are facing problems due to huge heap of data. The key is to find and develop new ways to efficiently process, manage and store the unstructured data rather than falling back on the "drinking from a fire hose" approach, in which a huge amount of data is coming to digital universe and very less data is analyzed properly. The issues of unstructured data lead to the challenges to extract useful information. Following are some of the unstructured data issues:

A. Data Quality and Usability Issues

Unstructured data has variety, scalability, speed, accuracy, and interactiveness issues and generating challenges in data analytics and mining [44]. Unstructured data has noise i.e. irrelevant or meaningless data [45]. Simplified access to data sources and absorbability of large amount of information may leads to increased amount of noise in data and decreases its quality [46]. Dirty data i.e. incomplete, inaccurate and improper data is a huge obstacle in IE. Unstructured data comes from different sources. So, it may contain redundant data with different representations, inaccurate data with false values, and inconsistent data with contradictions [47]. Separating dependable and solid data from unstructured data to determine enormous information is a basic issue that should be handled with ideal and right choices. Information vulnerability influences the quality of data. Advancements of models and techniques are required to extract reliable information from unstructured big data [48]. Variety of unstructured data makes it less trustworthy, and generating data quality and usability issues [49]. Removing this dirty data, both at individual data source level and integration of multiple sources level, is the ultimate need to efficiently manage, process and store unstructured data. Most of the unstructured data is unverified by nature. Poor quality of data leads to inaccurate and poor results at enterprise level which could be extremely costly. Some quality dimensions which are required to be considered to improve the results generated by unstructured data such as accuracy, complexity, completeness, usability, validity, and time factors.

B. Data Management Issues

Unstructured data management is a challenging task due to the scalability and complexity issues of unstructured data [50]. Unstructured data management is one of the major problems because unstructured data has no schema or pre-defined model. Advancement of unstructured data management systems, queries, more contextual search, content intelligence are important issues to be solved for

better management of unstructured data [51]. Inefficient accessibility of unstructured data, need of skilled persons and lack of expert knowledge are the challenges of big data management [52].

C. Heterogeneity Issues

Unstructured data is growing very fast and to extract useful information requires to precisely specify the tasks of IE process. Unstructured data has no structure and highly dynamic as compared to structured data. The complicated heterogeneity of mixed data makes it difficult to analyze and extract useful information [50] [53].

D. Data Variability Issues

High-performance simulations for computational chemistry, advanced and scalable algorithms for energy computation, interactive visualization systems, enhanced and efficient strategies for querying, and advanced data analytical tools are the utmost requirement in the field of computational sciences [54]. The evolution of smart industries is focused on the accurate and timely decision making from the real time data. Manufacturing intelligence improves operational efficiency, process innovation, and environmental impact. The success of manufacturing intelligence is hindered by the huge volume and complexity of unstructured data [55]. Emergence of e-Infrastructures in cross disciplinary collaboration with adequate governance model is a paradigm that can change science into e-science [56], but the complexity of data varieties is one of the challenges that has to be addressed prior to any successful implementation.

E. Feature Selection and Extraction issues

Feature extraction and transformation from unstructured data is more critical as compared to structured data due to the heterogeneity of unstructured documents. In this regard, a hybrid feature transformation techniques based on iterative classification with feature weighing have been proposed for multiple domains [57]. Although, feature transformation from heterogeneous unstructured data was achieved but minimal loss of precision was observed. Feature extraction and transformation needs advanced data preparation techniques. These techniques will help to improve the preprocessing and feature extraction tasks of heterogeneous, diverse and multi-dimensional unstructured data. Similarly, information extraction from unstructured clinical notes that contains inconsistent abbreviations and lack of structure can be achieved using matrix factorization, and multi-view learning technique in preprocessing and data modelling to handle the heterogeneous data [58]. While extracting features and preprocessing unstructured content, interpretability is an open quality dimension that should be considered.

IE from one type of data is different as compared to variety of data types. Preprocessing prior to IE from unstructured data will be helpful to resolve these issues. Mostly unstructured data comes from multiple (may be unreliable) sources. IE from unstructured data by understanding and avoiding the sources of errors is a challenging task. To the best of our knowledge, there is no single unified model to extract information from more than one type of data. For example, IE techniques for text in biomedical domain are not adequate to extract information from images of the same domain. In this regard, a mapping process for unstructured data in terms of data management has been explained in [59] where unstructured data are managed using DC metadata element 'subject'. But still the research is at very initial stage and facing scalability issues. Although metadata generation techniques can improve usability and scalability of unstructured data [60]. A multistep pipeline with data preprocessing steps, extraction methods and representation are utmost requirement to improve the unstructured data analytics. Problems with unstructured data are adding more challenges to the phases of the pipeline.

IV. CONCLUSION

Big data services provide a platform that could be emerged in any field of science and engineering to improve the productivity and development. In this paper, IE process for unstructured big data along with various methods and their corresponding technologies to transform unstructured data into useful information from diverse domains has been reviewed. In each method for extraction from various data types, latest techniques, trends, as well as challenges, have been elaborated. The variety of input data types are explored, and output is generated that represents the unstructured data into useful manner. As unstructured data growth rate is very high, the main concern of big data analytics is to extract useful structured information from unstructured big data. It is also very challenging to combine domain-specific and domain-independent solutions. The problems of unstructured data make IE process more challenging. Advanced preprocessing techniques prior to IE from unstructured data will be helpful to resolve these issues. To the best of our knowledge, there is no unified model to extract useful information from unstructured big data to improve the availability, quality and usability of unstructured data. Moreover, big data analytics with variety of data from diverse sources is facing many issues of quality and usability, data management, heterogeneity, and variability. Unstructured data is an important asset for organizations. However, to improve big data analytics, multistep pipeline for IE from multifaceted unstructured big data is an utmost requirement with data preprocessing steps, extraction methods and representation phases.

V. ACKNOWLEDGMENT

This research is funded by Universiti Tunku Abdul Rahman (UTAR) under the UTAR Research Fund (UTARRF): IPSR/RMC/UTARRF/2017-C1/R02.

REFERENCES

1. V. Turner, J. F. Gantz, D. Reinsel, and S. Minton, "The Digital Universe of Opportunities: Rich Data and Increasing Value of the Internet of Things," IDC White Paper, no. April, pp. 1–5, 2014.
2. M. T. Maybury, Multimedia information extraction : advances in video, audio, and imagery analysis for search, data mining, surveillance, and authoring. Wiley, 2012.
3. W. Zhu, P. Cui, Z. Wang, and G. Hua, "Multimedia Big Data Computing," IEEE MultiMedia, vol. 22, no. 3, pp. 96–105, Jul. 2015.
4. J. Piskorski and R. Yangarber, "Information Extraction: Past, Present and Future," in Multi-Source, Multilingual Information Extraction and Summerization, Springer, Berlin, Heidelberg, 2013, pp. 23–49.

5. N. Kanya and T. Ravi, "Modelings and techniques in named entity recognition: an information extraction task," in IET Chennai 3rd International Conference on Sustainable Energy and Intelligent Systems (SEISCON 2012), 2012, pp. 104–108.
6. A. Smirnova and P. Cudré-Mauroux, "Relation Extraction Using Distant Supervision," ACM Computing Surveys, vol. 51, no. 5, pp. 1–35, Nov. 2018.
7. L. Xue, S. Qing, and Z. Pengzhou, "Relation Extraction Based on Deep Learning," in 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018, pp. 687–691.
8. A. Ben Abacha and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," Information Processing & Management, vol. 51, no. 5, pp. 570–594, Sep. 2015.
9. A. Singhal, M. Simmons, and Z. Lu, "Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature," Journal of the American Medical Informatics Association, vol. 23, no. 4, pp. 766–772, Jul. 2016.
10. G. Zhou, L. Qian, and J. Fan, "Tree kernel-based semantic relation extraction with rich syntactic and semantic information," Information Sciences, vol. 180, no. 8, pp. 1313–1325, Apr. 2010.
11. S. Oramas, L. Espinosa-Anke, M. Sordo, H. Saggion, and X. Serra, "Information extraction for knowledge base construction in the music domain," Data & Knowledge Engineering, vol. 106, pp. 70–83, Nov. 2016.
12. P.-M. Ryu, M.-G. Jang, and H.-K. Kim, "Open domain question answering using Wikipedia-based knowledge model," Information Processing & Management, vol. 50, no. 5, pp. 683–692, Sep. 2014.
13. A. Dash, M. Pandey, and S. Rautaray, "Enhanced Entity Extraction Using Big Data Mechanics," in Advances in Intelligent Systems and Computing, Springer, Singapore, 2019, pp. 57–67.
14. F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong, "An Overview of Event Extraction from Text," in Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011), 2011, pp. 48–57.
15. C. Wang, X. Ma, J. Chen, and J. Chen, "Information extraction and knowledge graph construction from geoscience literature," Computers & Geosciences, vol. 112, pp. 112–120, Mar. 2018.
16. S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu---A Resource-Poor Language," ACM Transactions on Asian Language Information Processing, vol. 9, no. 4, pp. 1–43, Dec. 2010.
17. S. S. Sazali, N. A. Rahman, and Z. A. Bakar, "Information extraction: Evaluating named entity recognition from classical Malay documents," in 2016 Third International Conference on Information Retrieval and Knowledge Management (CAMP), 2016, pp. 48–53.
18. S. Ahmadi and Sina, "A Rule-Based Kurdish Text Transliteration System," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 18, no. 2, pp. 1–8, Jan. 2019.
19. J. J. Berman and J. J. Berman, "Chapter 1 – Providing Structure to Unstructured Data," in Principles of big data : preparing, sharing, and analyzing complex information, Morgan Kaufmann 2013, 2013, pp. 1–14.
20. W. Zhu, L. Luo, C. Ju, and B. Zhang, "Cross language information extraction for digitized textbooks of specific domains," in 2012 IEEE 12th International Conference on Computer and Information Technology, CIT 2012, 2012, pp. 1114–1118.
21. O. Rusu et al., "Converting unstructured and semi-structured data into knowledge," in 11th RoEduNet IEEE International Conference, 2013, pp. 1–4.
22. A. McCallum, "Information extraction: Distilling Structured Data from Unstructured Text," Queue - Social Computing, vol. 3, no. 9, pp. 48–57, 2005.
23. A. Agrawal, P. Mangalraj, and M. A. Bisherwal, "Target detection in SAR images using SIFT," in 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2015, pp. 90–94.
24. D. Ping Tian, "A Review on Image Feature Extraction and Representation Techniques," International Journal of Multimedia and Ubiquitous Engineering, vol. 8, no. 4, pp. 385–396, 2013.
25. C. Sumathi, T. Santhanam, and Gg. Devi, "A SURVEY ON VARIOUS APPROACHES OF TEXT EXTRACTION IN IMAGES," International Journal of Computer Science & Engineering Survey (IJCSES), vol. 3, no. 4, p. 27, 2012.
26. S. Deivalakshmi, R. Poreddy, P. Palanisamy, and S. Malakar, "Information extraction and unfilled-form structure retrieval from filled-up forms," in 2013 International Conference on Recent Trends in Information Technology (ICRTIT), 2013, pp. 297–300.
27. S. Aarthi and S. Chitrakala, "Scene understanding — A survey," in 2017 International Conference on Computer, Communication and Signal Processing (ICCCSP), 2017, pp. 1–4.
28. U. Sharma, S. Maheshkar, and A. N. Mishra, "Study of Robust Feature Extraction Techniques for Speech Recognition System," in 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), 2015, no. Ablaze, pp. 2–6.
29. U. Shrawankar and V. M. Thakare, "Techniques for Feature Extraction In Speech Recognition System : A Comparative Study," May 2013.
30. A. Kumar and B. Raj, "Audio Event Detection using Weakly Labeled Data," in Proceedings of the 24th ACM international conference on Multimedia - MM '16, 2016, pp. 1038–1047.
31. S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound Event Detection in Multichannel Audio Using Spatial and Harmonic Features," Jun. 2017.
32. E. Quinton, M. Sandler, and C. Harte, "EXTRACTION OF METRICAL STRUCTURE FROM MUSIC RECORDINGS," in Proceedings of the 18th International Conference on Digital Audio Effects (DAFx), 2015.
33. M. S. Hossain and G. Muhammad, "Healthcare Big Data Voice Pathology Assessment Framework," IEEE Access, vol. 4, pp. 7806–7815, 2016.
34. S. Hailemariam and K. Prahallad, "Extraction of Linguistic Information with the AID of Acoustic Data to Build Speech Systems," in International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, 2007, p. IV-717-IV-720.
35. D. Y. Mohammed, P. J. Duncan, M. M. Al-Maathidi, and F. F. Li, "A system for semantic information extraction from mixed soundtracks deploying MARSYAS framework," in 2015 IEEE 13th International Conference on Industrial Informatics (INDIN), 2015, pp. 1084–1089.
36. C.-H. Lee and S. M. Siniscalchi, "An Information-Extraction Approach to Speech Processing: Analysis, Detection, Verification, and Recognition," Proceedings of the IEEE, vol. 101, no. 5, pp. 1089–1115, May 2013.
37. Yi-Chong Zeng, "Automatic extraction of useful scenario information for dramatic videos," in 2013 9th International Conference on Information, Communications & Signal Processing, 2013, pp. 1–5.
38. Y.-S. Lee, C.-Y. Hsu, P.-C. Lin, C.-Y. Chen, and J.-C. Wang, "Video summarization based on face recognition and speaker verification," in 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), 2015, pp. 1821–1824.
39. Z. Zhang and H. Shi, "No-reference video quality assessment based on temporal information extraction," in 2nd International Symposium on Instrumentation and Measurement, Sensor Network and Automation (IMSNA), 2013, pp. 925–927.
40. B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous Knowledge Transfer in Video Emotion Recognition, Attribution and Summarization," IEEE Transactions on Affective Computing, vol. 9, no. 2, pp. 255–270, Apr. 2018.
41. J.-H. Park and K. Cho, "Extraction of visual information in basketball broadcasting video for event segmentation system," in 2016 International Conference on Information and Communication Technology Convergence (ICTC), 2016, pp. 1098–1100.
42. A. Mathur, T. Saxena, and R. Krishnamurthi, "Generating Subtitles Automatically Using Audio Extraction and Speech Recognition," in IEEE International Conference on Computational Intelligence & Communication Technology, 2015, pp. 621–626.
43. A. Biswas, P. K. Sahu, and M. Chandra, "Multiple camera in car audio–visual speech recognition using phonetic and visemic information," Computers & Electrical Engineering, vol. 47, pp. 35–50, Oct. 2015.
44. D. Che, M. Safran, and Z. Peng, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities," in Database Systems for Advanced Applications, Springer, Berlin, Heidelberg, 2013, pp. 1–15.
45. H. Xiong and M. Steinbach, "Enhancing data analysis with noise removal," IEEE Transactions on Knowledge and Data Engineering, vol. 18, no. 3, pp. 304–319, 2006.
46. EYGM, W. Ke, and T. Peng, "Big data Changing the way businesses," International Journal of Simulation: Systems, Science and Technology, vol. 16, no. April, p. 28, 2014.
47. P. Vashisht and V. Gupta, "Big data analytics techniques: A survey," in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 2015, pp. 264–269.

48. S. K. Singh, N. Mani, and B. Singh, "A Framework for Extracting Reliable Information from Unstructured Uncertain Big Data," in Intelligent Decision Technologies, Springer, Cham, 2016, pp. 175–185.
49. J. Gao and A. Koronios, "Unlock the Value of Unstructured Data in EAM," in Proceedings of the 7th World Congress on Engineering Asset Management (WCEAM), 2014, pp. 265–275.
50. K. U. Jaseena and J. M. David, "Issues, challenges, and solutions: Big data mining," in Computer Science & Information Technology (Computer Science Conference Proceeding CSCP), 2014, pp. 131–140.
51. R. Blumberg and S. Atre, "The problem with unstructured data," Dm Review, vol. 13, no. 42–49, p. 62, 2003.
52. Z. M. Hanapiyah, W. N. W. Hanafi, and S. Daud, "Issues, Challenges and Opportunities of Big Data Management in Higher Education Institutions in Malaysia," Indian Journal of Science and Technology, vol. 11, no. 4, 2018.
53. Alexandros Labrinidis, H. V. Jagadish, A. Labrinidis, and H. V. H. Jagadish, "Challenges and opportunities with big data," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2032–2033, Aug. 2012.
54. V. Yeguas and R. Casado, "Big Data issues in Computational Chemistry," in 2014 2nd International Conference on Future Internet of Things and Cloud (FiCloud), 2014, pp. 389–392.
55. P. O'Donovan, K. Leahy, K. Bruton, and D. T. J. O'Sullivan, "Big data in manufacturing: a systematic mapping study," Journal of Big Data, vol. 2, no. 1, p. 20, 2015.
56. Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in Collaboration Technologies and Systems (CTS), 2013 International Conference on, 2013, pp. 48–55.
57. R. K., H. Srinivas, and S. S., "Industrial information extraction through multi-phase classification using ontology for unstructured documents," Computers in Industry, vol. 100, pp. 137–147, Sep. 2018.
58. V. Huddar, B. K. Desiraju, V. Rajan, S. Bhattacharya, S. Roy, and C. K. Reddy, "Predicting Complications in Critical Care Using Heterogeneous Clinical Data," IEEE Access, vol. 4, pp. 7988–8001, 2016.
59. M. F. Abdullah and K. Ahmad, "The mapping process of unstructured data to structured data," in 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), 2013, pp. 151–155.
60. S. G. Small and L. Medsker, "Review of information extraction technologies and applications," Neural Computing and Applications, vol. 25, no. 3–4, pp. 533–548, Sep. 2014.

AUTHORS PROFILE

K. ADNAN received the B.S degree in Computer Science from Government College University Lahore Pakistan and M.S degree in Computer Science from Lahore College for Women University Lahore Pakistan in 2011. She is currently perusing the Ph.D. in Computer Science at Universiti Tunku Abdul Rahman Kampar Malaysia. From 2012 to 2016, she worked as Lecturer in Lahore Garrison University Lahore Pakistan. Currently she is a Research Assistant in UTAR (Universiti Tunku Abdul Rahman) Kampar Campus Malaysia. Her research interest includes Big Data processing, Information Extraction, Analytics and Data-driven Decision Making.

R. AKBAR received the MSc degree in Computer Science from University of Agriculture, Faisalabad Pakistan in 2001, MS degree in Computer Science with specialization in software engineering from Government College University, Lahore Pakistan in 2008 and the PhD degree in information technology from Universiti Teknologi, PETRONAS Malaysia, in 2013.

From 2001 to 2008, he was working as Lecturer in GC University, Lahore Pakistan. Later, he joined IT industry as Project Manager. Since 2012, he has been working with Faculty of Information and Communication Technology, Universiti Tunku Abdul Rahman, Malaysia as Lecturer and then Assistant Professor. His research interests include software processes, project behaviors, agile methodologies, process tailoring, big data and information security. He is reviewer of different journals and conferences.

Dr. Akbar has received teaching excellence award for year 2016.

K. S. WANG received the B.S. degree in Computer Science from Campbell University, USA and the M.S. degree in Advanced Information Technology from Universiti Malaysia Sarawak, Malaysia. Dr. Khor Siak Wang received his Ph.D. degree from Universiti Putra Malaysia (UPM) Malaysia in Information Retrieval.

He is currently working as Assistant Professor and Head of Postgraduate Programme in Department of Information Systems in Universiti Tunku Abdul Rahman Kampar campus Malaysia.