ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(17), DOI: 10.17485/ijst/2016/v9i17/93028, May 2016* ISSN (Online) : 0974-5645

Survey Big Data Analytics, Applications and Privacy Concerns

Ganesh D. Puri* **and D. Haritha**

CSE, KL University Vaddeswaram, Guntur Dist, Andhra Pradesh India;
puriganeshengg@gmail.com, haritha_donavalli@kluniversity.in

Abstract

Background/Objectives: The sources of big data are social media, enterprise data, unstructured data, sensor and clickstream data. The objective is to integrate this variety of data at one platform for processing the big data and find privacy concerns. **Methods:** The privacy concerns are raised due to unauthorized data extraction, collection and sharing information about user. For integrating and processing of big data; different tools and techniques are available. **Findings:** General framework for privacy preserving is discussed. Advancements in the big data analytics methods have posed different challenges in front of user. Due to large volume and variety of big data many organizations cannot process the data and needs to outsource it. While sharing such data for processing; there is need to apply proper privacy preserving measures. **Application/Improvements:** Privacy preserving techniques have applications in electronic health record processing, government surveys, outsourcing enterprise data for processing.

Keywords: Big Data, Big Data Analytics, Privacy Concerns, Privacy Preserving Methods

1. Introduction

Due to advancement in microprocessor electronics and availability of high performance communication networks abundant information is available. The data is getting generated in large quantity from number of sources. Data generation is estimated up to 2.5 Exabyte (1Exabyte=1,000,000 Terabytes) of data per day[1]. Figure 1 shows the exponential growth of the data. The sources for the data can be categorized in internal and external sources broadly. Figure 2 shows different sources of big data. The internal sources are application log, machine generated data, click stream data, sensor data etc. External sources of big data generation are social media, enterprise data such as transactions, emails, contracts. It also includes weather data, sensor generated data for vehicle, traffic, cell phone GPS signals. New York stock exchange generate 1 TB data, twitter generates 10TB data every day. This can be fed to sentiment analysis and based on this it can be discovered what people feel about various products and events. Volume is important to consider for example power meters are generating billons of reading every year and it is necessary to analyze this data to optimize the energy and actually see usage at energy per man.

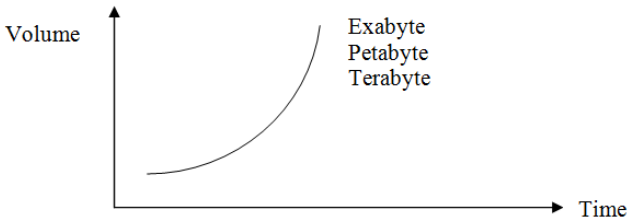

Figure 1. Big data generation at exponential growth.

Velocity is another important characteristic, for various time sensitive activities. For example to decide fraud, seconds can be decisive in being successful or not. The other aspect of the velocity is that the combing of data which is real time with the default should be possible. In other example A modern car is having 100 s sensors and sensors generating large volume of data arriving in very rapid way.95% of the data is being generated in unstructured or semi structured format[2]. As the population is increasing this uses smart phones. Business generate transaction data, but now users being

* *Author for correspondence*

on the internet generate tremendous amount of data images, videos text and it is need to process all of them. The number of smart phone users is increased to 75% up from 35% in 2011 in United States[3]. This availability of mobile devices made many things come into reality. The large content of information is available. The people get connected with others for communication virtually.

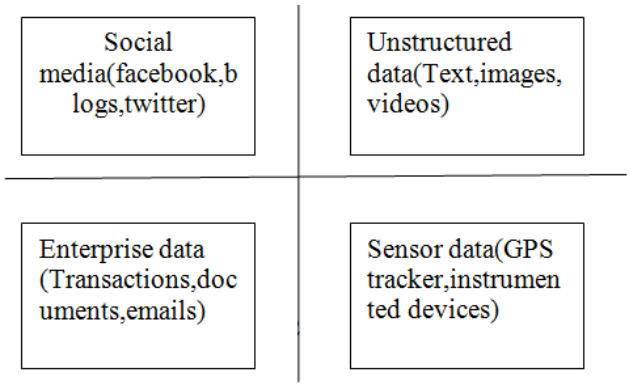

Figure 2. Big data sources.

2. Big Data Definitions

Data generating from different sources have different characteristics. In the definitions available in the literature are focusing on the large volume, variety, velocity of the data. The emphasis is on the processing capabilities or infrastructure availabilities available for processing, otherwise which was impossible with traditional framework.

The definitions for big data are leveraging on the ability of business intelligence, competitive intelligence, enhanced insight and decision making. In 2001 definition given by Laney[4] as :"high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". In 2012 the definition is updated in[5] as "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". The above definitions are emphasizing on 3V model i.e. high volume, high variety and high velocity. Few organizations added the term value in definition to make it 4V model. Afterward veracity term is added for big data to call it 5V model. In[6] addition of ambiguity, viscosity, and virality to the 3V model is discussed. Lack of metadata causes the ambiguity for example in the large volume of data M and F can be taken for March and February instead of male and female. Viscosity is the measure of resistance. Viscosity for example resistance in data flow, business rules and technology may cause loss of business. Virality measures how fast data can spread. For example re-tweets on a tweet. The ambiguity, viscosity and virality characteristics are useful from the point of analysis.

From the point of scalability to big data analytics the definition are suggested in[7] as attributive definition and architectural definition.

In attributive definition it says that according to a 2011 report that was sponsored by EMC (the cloud computing leader)[8]: Big data technologies require new platforms to store and process the data and derive the value from large volume and different forms of data.

In architectural definition the National Institute of Standards and Technology (NIST)[9] suggests that, Due to limitations of traditional relational approaches, processing of big data in large volume and variety of data which is coming at varying velocity, the need of scalability in the processing is required.

3. Big Data Processing

In processing of big data we have to consider diversity of data. The data is taken to one platform. Based on the internal and external sources steps can be identified as 1) acquisition of data from different sources, 2) processing 3) visualize 4) intelligence (Figure 3)

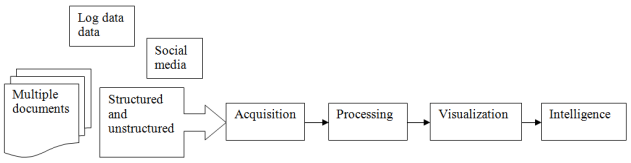

Figure 3. Big data processing flow.

3.1 Acquisition

Acquisition: data from different sources such as social media, application logs, clickstreams, emails, documents, SMS and phone calls are aggregated. Data integration tools can be helpful to integrate structured and unstructured data.

- Flume is distributed reliable tool for efficiently collecting and aggregating the log data. It works on streaming data flows[10]. It is fault tolerant and reliable

and support real-time analytic[6].

- Sqoop: For acquisition and integration of the data from RDBMS platform to hadoop platform for processing sqoop connector can be used. Enterprise data like transactions, metadata, data warehouse, data from enterprise system is taken to hadoop platform and processed in batch[10,6].

3.2 Processing

Processing: In data processing data collected in large volume is processed. Variety of data present two basic processing types. First is batch processing to process large volume of recorded data in the form of file. Second is real time processing to process large volume of data in the form of stream. Such data in large volume is stored in node of clusters on hadoop distributed file system. HDFS is scalable, fault tolerant framework for storing data. It uses datanodes and namenodes to provide the reliability using replication of data among distributed node in cluster.

3.2.1 Batch Processing

In the batch-processing paradigm, data are first stored and then analyzed[11].

- Mapreduce technology split up the large volume input file into chunks. These chunks are processed on different nodes in distributed fashion. Map function converts the chunk file into (key, value) pair format. Then all chunks output is shuffled and reduce function will reduce using the key[12]. For this purpose mapreduce use jobtracker and tasktracker nodes. There are number of tools used on this framework.
- Chukwa is opensource, data collection system built on HDFS and mapreduce framework. Chukwa deals with log files from web server. It also does analysis of collected data and has ability to display[13].
- Pig platform provides ability for large data analysis. Pig process data on parallel framework so it has ability to process large volume of data.
- Oozie is job scheduling system can be built up on HDFS. It is used to schedule flow of work. It is done by collection of control flow nodes. These nodes are used to denote the beginning and end of workflow. Action nodes are triggering the processing of workflow. The action can be mapreduce, HDFS operations, SSH, email[6,14].

3.2.2 Real Time Processing Systems

- Spark streaming: it is built based on spark and highly fault tolerant stream processing system. It performs batch processing computation on very small time period. It uses immutability feature of resilient distributed databases. D-stream[15] is the abstraction provided by spark streaming. Instead of continuous input the discretized input can be taken from RDD.
- S4 is simple scalable streaming system[16]. Because of its decentralized nature, messages are used as basic primitive for communication among the nodes of cluster. It is distributed stream processing system and takes advantage of message passing interface. Computation is done in distributed manner by dividing the input stream into different parts. This is done by series of processing elements. Automatic fault tolerance is not provided by S4. User has to make arrangement to keep the data and messages intact if one of the processing elements failed[17].

3.3 Visualization

It helps to get 360 degree view of social issue. Visualization is useful to draw the inferences and test the hypotheses. Javascripts and different open source tools are used to visualize the response of followers in case of social media. Authors in[18] showed emotions of viewers can be expressed on twitter and changes on incident. Joy, sadness and neutral views can be visualized. In case of reality shows to find the impact of show on national and global level. Understand the views of audience and summarize and represent in understandable format.

3.4 Intelligence

Enterprise top management can take smart decision from the visualization and patterns com out of big data analysis. For customer sentiment analysis can be helpful for marketing and product development. Email analysis is useful to target key customers and their perceptions. Customer reviews can be analyzed to find satisfaction of customers. Attrition modeling helps to understand mood of customer and take the moves in business. Response modeling is similar to attrition modeling. By predicting a negative behavior of customer the corrective actions can be taken for purchase or response.

4. Big Data Analytics

Based on the data the analytics technique is applied to make the inferences. Table 1 shows different techniques available in the text analysis. It includes text mining, data mining, machine learning, information retrieval, and natural language processing and sentiment analysis. As big data comprised of images, audio, video the techniques for audio analysis and video analysis are shown. The applications of the text analytics include Stock market prediction, healthcare, finance marketing, education, political, social sciences.

In social media analytics in content based analytics content filtering, ranking and tagging is done. Quick insight from existing database is possible. Using structure based analytics; analysis of large data over billion of records is possible. Using social graph and graph analytics identification of most influential accounts is done. Using activity graph identification of strong connectedness from large records is done. After finding such most influential people from the graph analytics from social media, special offers can be designed to those customers.

Audio analytics use transcription based and phonetic based approaches to analyze the audio contents. The application of this is customer care analysis and satisfaction analytics. Video analytics applications include automated security and surveillance systems. It also includes the application in retail industry. By observing the videos from customers interaction in supermarket the items can be placed.

Big data analytics has application in variety of areas. In real time monitoring of businesses it plays important role. To run competitive business and respond to continuously changing business environment, real time big data

Table 1. Big data analytics and applications

Paper	Type	Technique	Subtasks	Advantage	Application
[2,3,19,20]	Text analytics	Information extraction	Entity recognition Relation extraction	Evidence-based decision-making	Stock market Prediction
		Text summarization	a.extractive approach (1.identify main units in text and relationship in them 2. location and frequency of text) b.abstractive approach.(extract semantic info)	Report writing	scientific and news articles, advertisements, emails, and blogs.
		Question answering	1.IR based approach 2. knowledge based approach 3.Hybrid approach	Reduction in response time	healthcare, finance, marketing, and education
[1,3,21]	Text analytics	Sentiment analysis	document-level, sentence-level, and aspect-based	Finding positive or negative emotions	Marketing, finance, and the political and social sciences
[22,3,12,18]	Social media analytics	Content based analytics Structure based analytics	In structure base analytics 1) social graphs 2)activity graphs	Community detection Social influence analysis, Link prediction	Quick insights into the public perception 360 degree view of the social issue Ex. Facebook's "People You May Know" YouTube's "Recommended for You",
[3]	Audio analytics		1.transcript-based approach 2. phonetic based approach	Feedback from customers or agents To handle frustrated callers	customer call centers and healthcare
[3]	Video analytics		Server-based architecture Edge-based architecture	Placement of items	automated security and surveillance systems, retail industry

analytics is required. Highly transactional businesses produce vast amount of event data that can be managed by the cloud based architecture, which can process big data in real time[23].

In[12] applications based on social big data are considered. The social big data applications are divided in social big data applications related to marketing area, crime analysis area, health care area and user experiences based visualization. In[18] social media twitter is used to find emotions of the users based on the tweet. In this sentiment analysis is used to find the emotions.

In[19] content analysis is used to find the environmental disaster situations in the news paper archives. It describes the system which takes the archives of the newspapers as input and generates useful event summaries from unstructured text. It extracts geographic positions for the event and store in online database that can be searched and visualized using an interactive map. In[20] tweet analysis of academic libraries is done. The most frequently occurred words, bigrams, trigrams are found using text mining methods. Text mining and data mining methods are used to understand importance of social data in academic libraries to help in decision making and strategic planning. Big data analytics applications in[21] include Marketing, finance, and the political and social sciences.

Growing popularity and development in the big data analytics has provided advantage in many of the applications. The applications include retail industry, telecom industry, finance sector, medical diagnosis, banking, manufacturing etc. It provides the excellent result in big data analytics, by processing on large volume of data. At the same time the privacy concern about user is increased. In data mining, emerging topic is privacy preserving data mining. In recent years lot of research is undertaken on this area. PPDM is all about reducing the risk of data mining operations. It focuses on avoiding unwanted disclosure of the sensitive information in the different operations of knowledge data discovery. The operations include data collecting, preprocessing, publishing and information delivering. The aim of the PPDM is to protect the information for secondary usage by unsanctioned disclosure. But at the same time utility of the data should be intact after applying privacy preserving techniques. While applying the PPDM techniques sensitive information should not be used directly. In the mining if the results are of sensitive data, it should be excluded.

5. Big Data Challenges

5.1 Challenge 1

To make the business and personalized service the data is collected but which is unknowingly breaching the privacy of the people. For ex. In retail industry in a mart the collection of videos where customer has spent lot of time, which objects are handled by the customer from this preference of the customer can be known[2]. Even detailed analysis of video and speech or audio of conversation captured while the family is purchasing in mart can be done. This is helpful for the retailer for making the preference model, next best offer, discounts and placement of the products etc..

Challenge:- Such analysis can raise the privacy concerns also.

5.2 Challenge 2

Same is the case about data generated in terms of videos. The use of CCTV for security has increased the need of analysis of video contents. The videos generated and shared among the groups or individuals on social media have increased quantity of the data. On some social media sites, video content uploading limit for users is increased up to 72 hours per minute[22]. The high resolution video content of one second is equivalent to 2000 text pages. The main problem is integrating this variety of data and management of this data. The extraction of useful information from such data sources is challenging task[24].

Challenge:- Such large volume of content requires need of scalability in storage systems.

5.3 Challenge 3

95% of the data is being generated in unstructured or semi structured format[2]. As the population is increasing this uses smart phones. According to[3] the number of smart phone users is increased to 75% up from 35% in 2011 in United States. This availability of mobile devices made many things come into reality. The large content of information is available. The people get connected with others for communication virtually.

Challenge:- This connectedness exposes their information to third parties also[25].

5.4 Challenge 4

Interesting characteristic of big data is veracity; can we trust the data that we have? It is interesting that Van Paul, business leader stated that about one third of big data available in the organization is not trustworthy.

Challenge:- So determining the data is truthful is very important challenge for big data.

5.5 Challenge 5

Data is available indifferent formats such as structured and unstructured. Much of the unstructured data includes word and excel sheets, messages, tweets, images, audio, video. Few contents of this information may be sensitive in nature[26].

Challenge:- In such data personally identifiable information and intellectual property right violation may take place.

6. Privacy Concerns in Big Data

The information extraction policies of organization have increased the concerns of users about their privacy. The terms user and consumer is used interchangeably in the privacy section. The abundant information coming from sensors, location trackers, GPS, clickstream, log data can be treated as big data. Capturing and sharing such information may be the concern of users. While collecting the user related data there are number of privacy pitfalls, considered in[27]. Privacy related data is extracted in social media. In[28] showed that it is possible to show or identify the location of user from the tweets made by user. The basic machine learning and geotagged information is used for that. Also[29] showed that from geotagged twitter information the geographic coordinates can be extracted and it can be extended upto city of user or zipcode of location. In[30] proposed that image and structural analysis combined with content analysis on geotagged photos with textual tags collected from flicker can be used for finding location. In[31] authors have considered likes and dislikes which shows interest on facebook can reveal information about hidden information like location, feelings, relationship status.

Considering the above points the private information or collected data from social media should be manipulated so that risks can be reduced. Due to such privacy concerns about data collected on social media, the users of social media are reluctant to give correct information. Such problem is called as blackhole[32].

Big data characteristics like volume, velocity and variety are related to privacy concerns. Large amount of data means the breach of security and violations in the privacy. This leads to dishonesty with the consumer. High velocity data means data coming from sensors, GPS, clickstream. For such data real time analysis is required. This analysis can be used for short term prediction[26]. The organizations which don't have capability to store the big data, such organizations cannot handle the volume, velocity and complexity of big data. This data is produced at certain time and need to be outsourced. The cloud service providers are providing scalable storage capability as per demand[33,34]. But at the same time the privacy constraints should be applied while handing over this data to cloud service providers. Variety characteristic of big data suggest that data comes in different formats such as csv, images, videos, instant messages, signals. This structured and unstructured information may contain personally identifiable information and intellectual property. Such information capturing and sharing may leads to privacy violations[26].

According to surveys many organizations lack of comprehensiveness for addressing security and privacy issues. As per the EMC sponsored study conducted by IDC, only one third of the businesses have made the distinction in big data from traditional non big data and adapted tools and management approaches accordingly. Still many organizations use traditional databases as the main tools of handling data.

The consumers have expressed deep concern about dishonesty among the businesses and misuse of personal information. So consumers are reluctant to give the correct information. Many consumers have taken actions such as turning off information collecting system such as location tracking feature. Consumers are opposing the secondary uses of the data collected for different use[35].

7. Privacy Preserving Methods

To apply the privacy preserving techniques we have to consider the different dimensions. In multidimensional dataset to find sensitive attributes, quasi identifiers and non sensitive attributes; different attribute selection methods should be applied[36]. These methods include Information Gain, Gain ratio, Pearson Correlation, Gini Index. After selection of key identifiers; these identifiers should be modified such that information will not be released to unauthorized user but at the same time utility of data will remain unchanged (Figure 4).

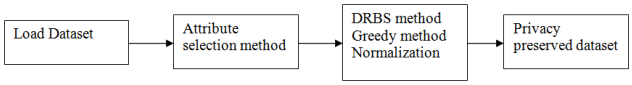

Figure 4. Privacy preserving flowgraph.

The methods available for perturbation of key identifiers are data relocation based sub clustering(DRBS)[37], Greedy method, Normalization[38]. In clustering based method; the clusters are found with centroid. Again clustering is applied to find subclusters. Then distance between the centroid of cluster and parent cluster is found and based on distance subclusters are arranged. The elements are rotated to neighbor cluster until last element is visited[39]. In normalization method for perturbation the key identifier values normalized.

8. Conclusion

In this paper need of big data processing is addressed. Advancement in big data analytics is useful for drawing inferences; at the same time it is main reason for increasing privacy concerns of user. Framework for privacy preserving is discussed.

9. References

1. IBM, Big Data and Analytics, 2015. Available from http://www01.ibm.com/software/data/bigdata/ what-is-big-data.html
2. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2015; 35:137–44.
3. December 2014 U.S. Smartphone subscriber market share. comScore. 2015. Available from https://www.comscore.com/Insights/Market-Rankings/comScore-Reports-December-2014-US-Smartphone-Subscriber-Market-Share.
4. Beyer MA, Laney D. The Importance of 'BigData': A Definition, Gartner, Stam-ford, CT; 2012.
5. Gartner IT Glossary (n.d.). Available from http://www.gartner.com/it-glossary/big-data/
6. Krishnan K. Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence. Elsevier Science; 2013.
7. Hu H, Wen Y, Tat-Seng C, Li X. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access, Practical Innovations: Open Solutions. 2014 Jul; 1–36.
8. Gantz J, Reinsel D. Extracting value from chaos. Proc. IDC iView. 2011; 1–12.
9. Cooper M, Mell P. Tackling Big Data. 2012. Available from http://csrc.nist.gov/groups/SMA/forum/documents/june-2012presentations/f%csm_june2012_cooper_mell.pdf
10. Wang C, Rayan IA, Schwan K. Faster, larger, easier: reining real-time big data processing in cloud. Proceedings of the Posters and Demo Track, Middleware '12, ACM; 2012; New York, NY, USA. pp. 4:1–4:2
11. Dean J, Ghemawat S. Mapreduce: Simplified data processing on large clusters. Commun. ACM. 2008; 51(1):107113.
12. Bello-Orgaz G, Jung JJ. David Camacho, Social big data: Recent achievements and new challenges. Information Fusion. 2016; 28:45–59.
13. Boulon J, Konwinski A, Qi R, Rabkin A, Yang E, Yang M. Chukwa, a large-scale monitoring system. Cloud Comput Appl. 2008; 1–5.
14. Emani CK, Cullot N, Nicolle C. Understandable Big Data: A survey. Computer science review. 2015; 17:70–81.
15. Zaharia M, Das T, Li H, Hunter T, Shenker S, Stoica I. Discretized streams: Fault-tolerant streaming computation atscale. Proc 24th ACM Symp Operating Syst Principles. 2013; 423–38.
16. Neumeyer L, Robbins B, Nair A, Kesari A. S4: Distributed stream computing platform. Proc IEEE Int Conf Data Mining Workshops; 2010. pp. 170–7.
17. Zhang H, Chen G, Ooi BC, Kian-Lee T, Zhang M. In-Memory Big Data Management and Processing: A Survey IEEE transactions on knowledge and data engineering. 2015 Jul; 27(7).
18. Yu Y, Wang X. World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fan's tweets. Computers in Human Behavior. 2015; 48:392–400.
19. Yzaguirre A, Warren R, Smit M. Detecting environmental disasters in digital news archives. IEEE International Conference on Big Data (Big Data); 2015. pp. 1–9.
20. Al-Daihani SM, Abrahams A. A Text Mining Analysis of Academic Libraries' Tweets. The Journal of Academic Librarianship. 2016. Available from http://dx.doi.org/10.1016/j.acalib.2015.12.014
21. Arora D, Malik P. Analytics: Key to go from generating big data to deriving business value. IEEE First International Conference on Big Data Computing Service and Applications; 2015. pp. 1–7.
22. Infographic. The Data Explosion in 2014 Minute by Minute. 2015. Available from http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic
23. Vera-Baquero A, Colomo-Palacios R, Molloy O. Real-time business activity monitoring and analysis of process performance on big-data domains. Telematics and Informatics. 2016; 33:793–807.
24. Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. IEEE Trans Knowl Data Eng. 2014; 26(1):97–107.
25. Eastin MS, Brinson NH, Doorey A, Wilcox G. Living in a big data world: Predicting mobile commerce activity through privacy concerns. Computers in Human Behavior. 2016; 58:214–20.
26. Kshetri N. Big data's impact on privacy, security and consumer welfare. Telecommunicatios Policy. 2014; 38:1134–45.
27. Ferrara E, De Meo P, Fiumara G. Robert Baumgartner, Web data extraction, applications and techniques: A survey. Knowledge-Based Systems. 2014; 70:301–23.
28. Hecht B, Hong L, Suh B, Chi E. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. Proc International Conference on Human Factors in Computing Systems; 2011; British Columbia, Canada. pp. 237–46.

29. Kinsella S, Murdock V, O'Hare N. I'm eating a sandwich in Glasgow: modeling locations with tweets. Proc International Workshop on Search and Mining User-generated Contents, ACM; 2011; Glasgow, UK. pp. 61–8.
30. Crandall D, Backstrom L, Huttenlocher D, Kleinberg J. Mapping the world's photos. Proc 18th International Conference on World Wide Web, ACM; 2009; Madrid, Spain. pp. 761–70.
31. Chaabane A, Acs G, Kaafar M. You are what you like! information leakage through users' interests. Proc Annual Network and Distributed System Security Symposium; 2012.
32. Ye S, Lang J, Wu F. Crawling Online Social Graphs. Proc. 12th International Asia-Pacific Web Conference; 2010. pp. 236–42. IEEE; 2010.
33. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU. The rise of "big data" on cloud computing:Review and open research issues. Information Systems. 2015; 47:98–115.
34. Andreolini M, Colajanni M, Pietri M, Tosi S. Adaptive, scalable and reliable monitoring of big data on clouds. J Parallel Distrib Comput. 2015; 67–80.
35. Mayer-Schönberger V, Cukier K. Big Data: A revolution that will transform how we live, work and think. Boston: Houghton Mifflin Harcourt; 2013.
36. Sudha M, Kumaravel A. Performance Comparison based on Attribute Selection Tools for Data Mining. Indian Journal of Science and Technology. 2014 Nov; 7(S7):1–5.
37. Rajalakshmi V, Mala AGS. Anonymization by Data Relocation Using Sub-clustering for Privacy Preserving Data Mining. Indian Journal of Science and Technology. 2014 Jan; 7(7):1–6.
38. Manikandan G, Sairam N, Sharmili S, Venkatakrishnan S. Achieving Privacy in Data Mining using Normalization. Indian Journal of Science and Technology. 2013 Apr; 6(4):1–5.
39. Hariharan R, Mahesh C, Prasenna P, Kumar RV. Enhancing Privacy Preservation in Data Mining using Cluster based Greedy Method in Hierarchical Approach. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–8.