# Paper Presentation on Data Mining for Internet Of Things

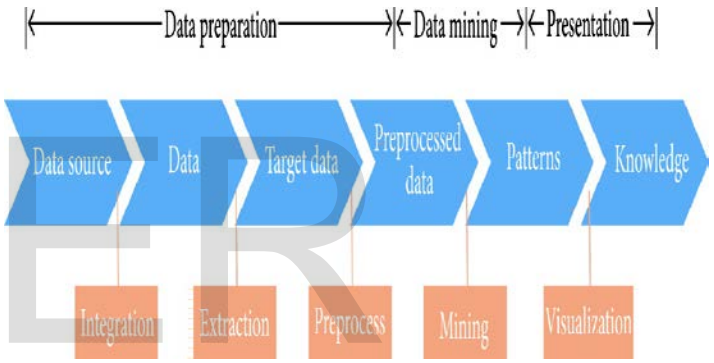
**Authors:** P.Chandana,G.Hemalatha

**Abstract**—The massive data generated by the Internet of Things (IoT) are considered of high business value, and data mining algorithms can be applied to IoT to extract hidden information from data. In this paper, we give a systematic way to review data mining in knowledge view, technique view, and application view, including classification, clustering, association analysis, time series analysis and outlier analysis. And the latest application cases are also surveyed. As more and more devices connected to IoT, large volume of data should be analyzed, the latest algorithms should be modified to apply to big data. We reviewed these algorithms and discussed challenges and open research issues. At last a suggested big data mining system is proposed.

**Index Terms**— Minimum 7 keywords are mandatory, Keywords should closely reflect the topic and should optimally characterize the paper. Use about four key words or phrases in alphabetical order, separated by commas.

## 1 INTRODUCTION –

The Internet of Things (IoT) and its relevant technologies can seamlessly integrate classical networks with networked instruments and devices. IoT has been playing an essential role ever since it appeared, which covers from traditional equipment to general household objects and has been attracting the attention of researchers from academia, industry, and government in recent years. There is a great vision that all things can be easily controlled and monitored, can be identified automatically by other things, can communicate with each other through internet, and can even make decisions by themselves. In order to make IoT smarter, lots of analysis technologies are introduced into IoT; one of the most valuable technologies is data mining. Data mining involves discovering novel, interesting, and potentially useful patterns from large data sets and applying algorithms to the extraction of hidden information. Many other terms are used for data mining, for example, knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, and information harvesting. The objective of any data mining process is to build an efficient predictive or descriptive model of a large amount of data that not only best fits or explains it, but is also able to generalize to new data. Based on a broad view of data mining functionality, data mining is the process of discovering interesting knowledge from large amounts of data stored in either databases, data warehouses, or other information repositories. On the basis of the definition of data mining and the definition of data mining functions, a typical data mining process includes the following steps.




i. **Data preparation:** prepare the data for mining. It includes 3 substeps: integrate data in various data sources and clean the noise from data; extract some parts of data into data mining system; preprocess the data to facilitate the data mining.

ii. **Data mining:** apply algorithms to the data to find the patterns and evaluate patterns of discovered knowledge.

iii. **Data presentation:** visualize the data and represent mined knowledge to the user.

We can view data mining in a multidimensional view.

(i) In knowledge view or data mining functions view, it includes characterization, discrimination, classification, clustering, association analysis, time series analysis, and outlier analysis.

(ii) In utilized techniques view, it includes machine learning, statistics, pattern recognition, big data, support vector machine, rough set, neural networks, and evolutionary algorithms.

(iii) In application view, it includes industry, telecommunication, banking, fraud analysis, biodata mining, stock market analysis, text mining, web mining, social network, and e-commerce.

A variety of researches focusing on knowledge view, technique view, and application view can be found in the literature. However, no previous effort has been made to review the different views of data mining in a systematic way, especially in nowadays big data; mobile internet and Internet of Things grow rapidly and some data mining researchers shift their attention from data mining to big data. There are lots of data that can be mined, for example, database data (relational database, NoSQL database), data warehouse, data stream, spatiotemporal, time series, sequence, text and web, multimedia, graphs, the World Wide Web, Internet of Things data, and *legacy* system log. Motivated by this, in this paper, we attempt to make a comprehensive survey of the important recent developments of data mining research. This survey focuses on knowledge view, utilized techniques view, and application view of data mining. Our main contribution in this paper is that we selected some wellknown algorithms and studied their strengths and limitations. The contribution of this paper includes 3 parts: the first part is that we propose a novel way to review data mining in knowledge view, technique view, and application view; the second part is that we discuss the new characteristics of big data and analyze the challenges. Another important contribution is that we propose a suggested big data mining system. It is valuable for readers if they want to construct a big data mining system with open source technologies. The rest of the paper is organized as follows. In Section 2 we survey the main data mining functions from knowledge view and technology view, including classification, clustering, association analysis, and outlier analysis, and introduce which techniques can support these functions. In Section 3 we review the data mining applications in ecommerce, industry, health care, and public service and discuss which knowledge and technology can be applied to these applications. In Section 4, IoT and big data are discussed comprehensively, the new technologies to mine big data for IoT are surveyed, the challenges in big data era are overviewed, and a new big data mining system architecture for IoT is proposed. In Section 5 we give a conclusion.

## 2. Data Mining Functionalities

Data mining functionalities include classification, clustering, association analysis, time series analysis, and outlier analysis.

(i) Classification is the process of finding a set of models or functions that describe and distinguish data classes or concepts, for the purpose of predicting the class of objects whose class label is unknown.

(ii) Clustering analyzes data objects without consulting a known class model.

(iii) Association analysis is the discovery of association rules displaying attribute-value conditions that frequently occur together in a given set of data.

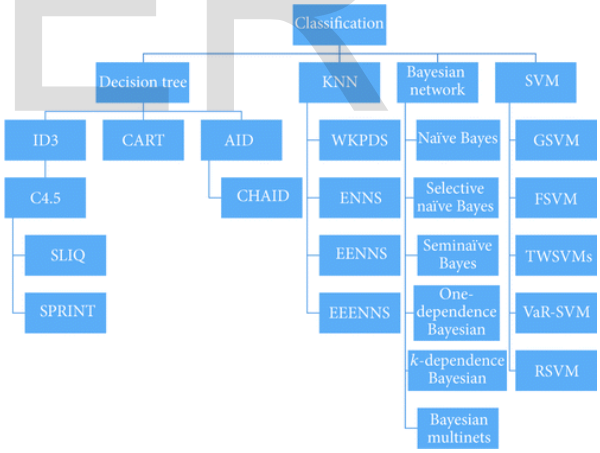
(iv) Time series analysis comprises methods and techniques for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

(v) Outlier analysis describes and models regularities or trends for objects whose behavior changes over time.

### 2.1. Classification.

Classification is important for management of decision making. Given an object, assigning it to one of predefined target categories or classes is called classification.The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

There are many methods to classify the data, including decision tree induction, frame-based or rule-based expert systems, hierarchical classification, neural networks, Bayesian network, and support vector machines.




(i) A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labeled with distinct outcomes of the test. Each leaf node has a class label associated with it. Iterative Dichotomiser 3 or ID3 is a simple decision tree learning algorithm. C4.5 algorithm is an improved version of ID3; it uses gain ratio as splitting criteria. The difference between ID3 and C4.5 algorithm is that

ID3 uses binary splits, whereas C4.5 algorithm uses multiway splits. SLIQ (Supervised Learning In Quest) is capable of handling large data sets with ease and lesser time complexity, SPRINT (Scalable Parallelizable Induction of Decision Tree algorithm) is also fast and highly scalable, and there is no storage constraint on larger data sets in SPRINT. Other improvement researches are finished. Classification and Regression Trees (CART) is a nonparametric decision tree algorithm. It produces either classification or regression trees, based on whether the response variable is categorical or continuous. CHAID (chi-squared automatic interaction detector) and the improvement researcher focus on dividing a data set into exclusive and exhaustive segments that differ with respect to the response variable.

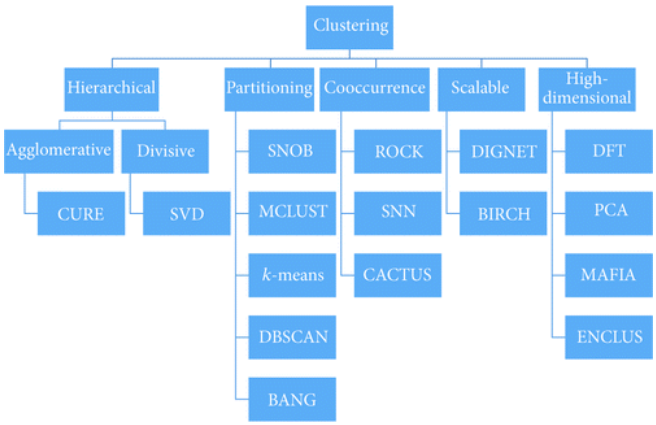
(ii) The KNN (K-Nearest Neighbor) algorithm is introduced by the Nearest Neighbor algorithm which is designed to find the nearest point of the observed object. The main idea of the KNN algorithm is to find the K-nearest points. There are a lot of different improvements for the traditional KNN algorithm, such as the Wavelet Based K-Nearest Neighbor Partial Distance Search(WKPDS) algorithm, EqualAverage Nearest Neighbor Search (ENNS) algorithm, Equal-Average Equal-Norm Nearest Neighbor code word Search (EENNS) algorithm, the Equal-Average Equal-Variance Equal-Norm Nearest Neighbor Search (EEENNS) algorithm, and other improvements.

(iii) Bayesian networks are directed acyclic graphs whose nodes represent random variables in the Bayesian sense. Edges represent conditional dependencies; nodes which are not connected represent variables which are conditionally independent of each other. Based on Bayesian networks, these classifiers have many strengths, like model interpretability and accommodation to complex data and classification problem settings. The research includes na¨ıve Bayes, selective na¨ıve Bayes, semina¨ıve Bayes, one-dependence Bayesian classifiers, K-dependence Bayesian classifiers , Bayesian network-augmented na¨ıve Bayes, unrestricted Bayesian classifiers, and Bayesian multinets.

(iv) Support Vector Machines algorithm is supervised learning model with associated learning algorithms that analyze data and recognize patterns, which is based on statistical learning theory. SVM produces a binary classifier, the so-called optimal separating hyperplanes, through an extremely nonlinear mapping of the input vectors into the high-dimensional feature space. SVM is widely used in text classification, marketing, pattern recognition, and medical diagnosis. A lot of further research is done, GSVM (granular support vector machines), FSVM (fuzzy support vector machines), TWSVMs (twin support vector machines), VaR-SVM (value-at-risk support vector machines), and RSVM (ranking support vector machines).

## 2.2. Clustering.

Clustering algorithms divide data into meaningful groups




so that patterns in the same group are similar in some sense and patterns in different group are dissimilar in the same sense. Searching for clusters involves unsupervised learning [56]. In information retrieval, for example, the search engine clusters billions of web pages into different groups, such as news, reviews, videos, and audios. One straightforward example of clustering problem is to divide points into different groups.

(i) Hierarchical clustering method combines data objects into subgroups; those subgroups merge into larger and high level groups and so forth and form a hierarchy tree. Hierarchical clustering methods have two classifications, agglomerative (bottom-up) and divisive (top-down) approaches. The agglomerative clustering starts with one-point clusters and recursively merges two or more of the clusters. The divisive clustering in contrast is a top-down strategy; it starts with a single cluster containing all data points and recursively splits that cluster into appropriate subclusters. CURE (Clustering Using Representatives) and SVD (Singular Value Decomposition) are typical research.

(ii) Partitioning algorithms discover clusters either by iteratively relocating points between subsets or by identifying areas heavily populated with data. The related research includes SNOB, MCLUST, k-medoids, and k-means related research. Density-based partitioning methods attempt to discover low-dimensional data, which is denseconnected, known as spatial data.The related research includes DBSCAN (Density Based Spatial Clustering of Applications

with Noise). Grid based partitioning algorithms use hierarchical agglomeration as one phase of processing and perform space segmentation and then aggregate appropriate segments; researches include BANG.

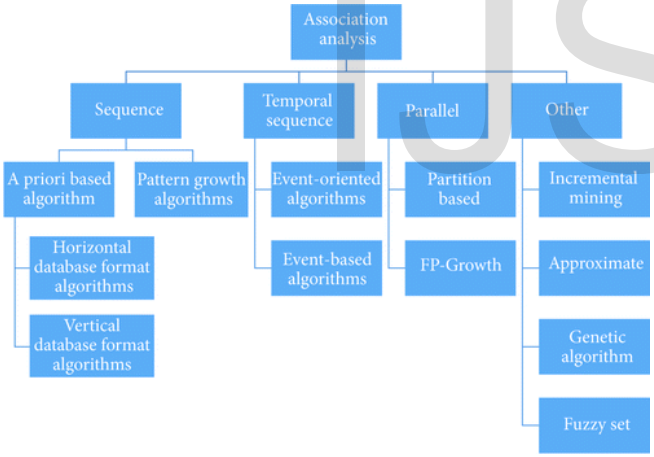
(iii) In order to handle categorical data, researchers change data clustering to preclustering of items or categorical attribute values; typical research includes ROCK.

(iv) Scalable clustering research faces scalability problems for computing time and memory requirements, including DIGNET and BIRCH.

(v) High dimensionality data clustering methods are designed to handle data with hundreds of attributes, including DFT and MAFIA.

### 2.3. Association Analysis.

Association rule mining focuses on the market basket analysis or transaction data analysis, and it targets discovery of rules showing attributevalue associations that occur frequently and also help in the generation of more general and qualitative knowledge which in turn helps in decision makin.



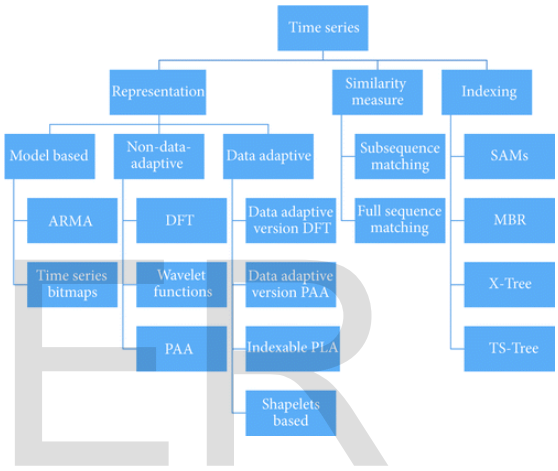

(i) For the first catalog of association analysis algorithms, the data will be processed sequentially. The a priori based algorithms have been used to discover intratransaction associations and then discover associations; there are lots of extension algorithms. According to the data record format, it clusters into 2 types: Horizontal Database Format Algorithms and Vertical Database Format Algorithms; the typical algorithms include MSPS and LAPIN-SPAM. Pattern growth algorithm is more complex but can be faster to calculate given large volumes of data. The typical algorithm is FP-Growth algorithm.

(ii) In some area, the data would be a flow of events and therefore the problem would be to discover event patterns that occur frequently together. It divides into 2 parts: event-based algorithms and event-oriented algorithms; the typical algorithm is PROWL.

(iii) In order to take advantage of distributed parallel computer systems, some algorithms are developed, for example, Par-CSP.

### 2.4. Time Series Analysis.

A time series is a collection of temporal data objects; the characteristics of time series data include large data size, high dimensionality, and updating continuously. Commonly, time series task relies on 3 parts of components, including representation, similarity measures, and indexing.




(i) One of the major reasons for time series representation is to reduce the dimension, and it divides into three categories:

model based representation,

nondata-adaptive representation,

and data adaptive representation.

The model based representations want to find parameters of underlying model for a representation.. In data adaptive representations, the parameters of a transformation will change according to the data available and related works including representations version of DFT /PAA and indexable PLA .

(ii) The similarity measure of time series analysis is typically carried out in an approximate manner; the research directions include subsequence matching and full sequence matching.

(iii) The indexing of time series analysis is closely associated with representation and similarity measure part; the research topic includes SAMs (Spatial Access Methods) and TS-Tree.

IJSER

### 2.5. Other Analysis.

Outlier detection refers to the problem of finding patterns in data that are very different from the rest of the data based on appropriate metrics. Such a pattern often contains useful information regarding abnormal behavior of the system described by the data. Distancebased algorithms calculate the distances among objects in the data with geometric interpretation. Density-based algorithms estimate the density distribution of the input space and then identify outliers as those lying in low density. Rough sets based algorithms introduce rough sets or fuzzy rough sets to identify outliers.

## 3. Data Mining Applications

### 3.1. Data Mining in e-Commerce.

Data mining enables the businesses to understand the patterns hidden inside past purchase transactions, thus helping in planning and launching new marketing campaigns in prompt and cost-effective way. e-commerce is one of the most prospective domains for data mining because data records, including customer data, product data, users' action log data, are plentiful; IT team has enriched data mining skill and return on investment can be measured. Researchers leverage association analysis and clustering to provide the insight of what product combinations were purchased; it encourages customers to purchase related products that they may have been missed or overlooked. Users' behaviors are monitored and analyzed to find similarities and patterns in Web surfing behavior so that the Web can be more successful in meeting user needs. A complementary method of identifying potentially interesting content uses data on the preference of a set of users, called collaborative filtering or recommender systems, and it leverages user's correlation and other similarity metrics to identify and cluster similar user profiles for the purpose of recommending informational items to users. And the recommender system also extends to social network, education area, academic library, and tourism.

### 3.2. Data Mining in Industry.

Data mining can highly benefit industries such as retail, banking, and telecommunications; classification and clustering can be applied to this area. One of the key success factors of insurance organizations and banks is the assessment of borrowers' credit worthiness in advance during the credit evaluation process. Credit scoring becomes more and more important and several data mining methods are applied for credit scoring problem. Retailers collect customer information, related transactions information, and product information to significantly improve accuracy of product demand forecasting, assortment optimization, product recommendation, and ranking across retailers and manufacturers. Researchers leverage SVM, support vector regression, or Bass model to forecast the products' demand.

### 3.3. Data Mining in Health Care.

In health care, data mining is becoming increasingly popular, if not increasingly essential. Heterogeneous medical data have been generated in various health care organizations, including payers, medicine providers, pharmaceuticals information, prescription information, doctor's notes, or clinical records produced day by day. These quantitative data can be used to do clinical text mining, predictive modeling, survival analysis, patient similarity analysis, and clustering, to improve care treatmentand reduce waste. In health care area, association analysis, clustering, and outlier analysis can be applied.

Treatment record data can be mined to explore ways to cut costs and deliver better medicine. Data mining also can be used to identify and understand high-cost patients and applied to mass of data generated by millions of prescriptions, operations, and treatment courses to identify unusual patterns and uncover fraud.

### 3.4. Data Mining in City Governance.

In public service area, data mining can be used to discover public needs and improve service performance, decision making with automated systems to decrease risks, classification, clustering, and time series analysis which can be developed to solve this area problem.

E-government improves quality of government service, cost savings, wider political participation, and more effective policies and programs, and it has also been proposed as a solution for increasing citizen communication with government agencies and, ultimately, political trust. City incident information management system can integrate data mining methods to provide a comprehensive assessment of the impact of natural disasters on the agricultural production and rank disaster affected areas objectively and assist governments in disaster preparation and resource allocation.

By using data analytics, researchers can predict which residents are likely to move away from the city, and it helps to infer which factors of city life and city services lead to a resident's decision to leave the city.

TABLE 1: The data mining application and most popular data mining functionalities.

| Application | Classification | Clustering | Association analysis | Time series analysis | Outlier analysis |
|---|---|---|---|---|---|
| e-commerce | | √ | √ | | |
| Industry | √ | √ | √ | | |
| Health care | | √ | √ | | √ |
| City governance | √ | √ | √ | √ | |

A major challenge for the government and lawenforcement is how to quickly analyze the growing volumes of crime data. Researchers introduce spatial data mining technique to find out the association rules between the crime hot spots and spatial landscape; other researchers leverage enhanced k-means clustering algorithm to discover crime patterns and use semisupervised learning technique for knowledge discovery and to help increase the

predictive accuracy. Also data mining can be used to detect criminal identity deceptions by analyzing people information such as name, address, date of birth, and social-security number and to uncover previously unknown structural patterns from criminal networks. In transport system, data mining can be used for map refinement according to GPS traces, and based on multiple users' GPS trajectories researchers discover the interesting locations and classical travel sequences for location recommendation and travel recommendation.

### 3.5. Summary.

The data mining application and most popular data mining functionalities can be summarized in Table 1.

## 4. Challenges and Open Research Issues in IoT and Big Data Era

With the rapid development of IoT, big data, and cloud computing, the most fundamental challenge is to explore the large volumes of data and extract useful information or knowledge for future actions. The key characteristics of the data in IoT era can be considered as big data; they are as follows.

(i) Large volumes of data to read and write: the amount of data can be TB (terabytes), even PB (petabytes) and ZB (zettabyte), so we need to explore fast and effective mechanisms.

(ii) Heterogeneous data sources and data types to integrate: in big data era, the data sources are diverse; for example, we need to integrate sensors data, cameras data, social media data, and so on and all these data are different in format, byte, binary, string, number, and so forth. We need to communicate with different types of devices and different systems and also need to extract data from web pages.

(iii) Complex knowledge to extract: the knowledge is deeply hidden in large volumes of data and the knowledge is not straightforward, so we need to analyze the properties of data and find the association of different data.

### 4.1. Challenges.

There are lots of challenges when IoT and big data come; the quantity of data is big but the quality is low and the data are various from different data sources inherently possessing a great many different types and representation forms, and the data is heterogeneous, as-structured, semistructured, and even entirely unstructured. We analyze the challenges in data extracting, data mining algorithms, and data mining system area. Challenges are summarized below.

(i) The first challenge is to access, extracting large scale data from different data storage locations. We need to deal with the variety, heterogeneity, and noise of the data, and it is a big challenge to find the fault and even harder to correct the data. In data mining algorithms area, how to modify traditional algorithms to big data environment is a big challenge.

(ii) Second challenge is how to mine uncertain and incomplete data for big data applications. In data mining system, an effective and security solution to share data between different applications and systems is one of the most important challenges, since sensitive information, such as banking transactions and medical records, should be a matter of concern.

### 4.2. Open Research Issues.

In big data era, there are some open research issues including data checking, parallel programming model, and big data mining framework.

(i) There are lots of researches on finding errors hidden in data, such as. Also the data cleaning, filtering, and reduction mechanisms are introduced.

(ii) Parallel programming model is introduced to data mining and some algorithms are adopted to be applied in it. Researchers have expanded existing data mining methods in many ways, including the efficiency improvement of single-source knowledge discovery methods, designing a data mining mechanism from a multisource perspective, and the study of dynamic data mining methods and the analysis of stream data. For example, parallel association rule mining and parallel k-means algorithm based on Hadoop platform are good practice. But there are still some algorithms which are not adapted to parallel platform, this constraint on applying data mining technology to big data platform. This would be a challenge for data mining related researchers and also a great direction.

(iii) The most important work for big data mining system is to develop an efficient framework to support big data mining. In the big data mining framework, we need to consider the security of data, the privacy, the data sharing mechanism, the growth of data size, and so forth. A well designed data mining framework for big data is a very important direction and a big challenge.

### 4.3. Recent Works of Big Data Mining System for IoT.

In data mining system area, many large companies as Facebook, Yahoo, and Twitter benefit and contribute works to opensource projects. Big data mining infrastructure includes the following.

(i) Apache Mahout project implements a wide range of machine learning and data mining algorithms.

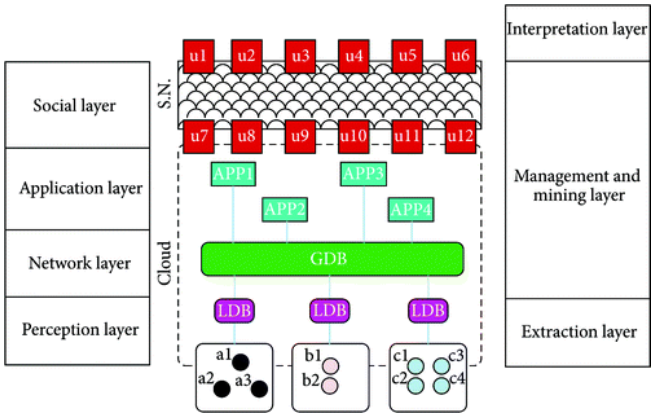
(ii) R Project is a programming language and software environment designed for statistical computing and visualization.

(iii) MOA project performs data mining in real time and SAMOA project integrates MOA with Strom and S4.

(iv) Pegasus is a petascale graph mining library for the Hadoop platform.

IJSER

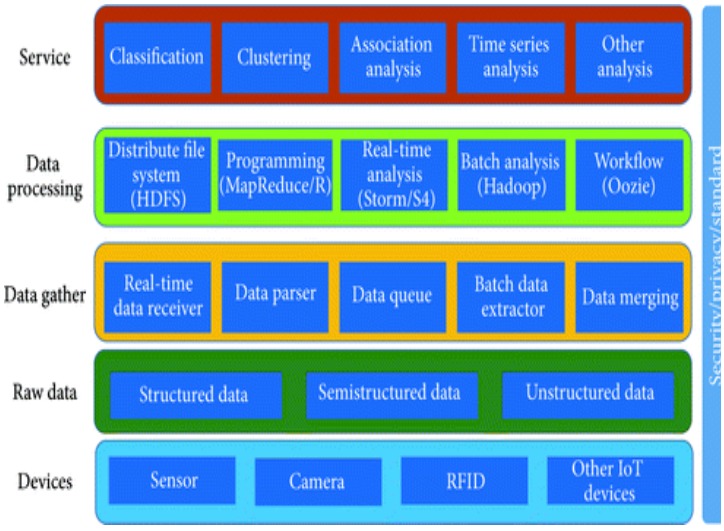
Some researchers from IoT area also proposed big data mining system architectures for IoT, and these systems focus on the integration with devices and data mining technologies. The following diagram shows an architecture for the support of social network and cloud computing in IoT.




They integrated the big data and KDD into the extraction, management and mining, and interpretation layers. The extraction layer maps onto the perception layer. Different from the traditional KDD, the extraction layer of the proposed framework also takes into consideration the behavior of agents for its devices.

### 4.4. Suggested System Architecture for IoT.

According to the survey of big data mining system and IoT system, we suggest the system architecture for IoT and big data mining system. In this system, it includes 5 layers.




**(i) Devices:**

lots of IoT devices, such as sensors, RFID, cameras, and other devices, can be integrated into this system to apperceive the world and generate data continuously.

**(ii) Raw data:**Iin the big data mining system, structured data, semistructured data, and unstructured data can be integrated.

**(iii) Data gather:** Real-time data and batch data can be supported and all data can be parsed, analyzed, and merged.

**(iv) Data processing:** Lots of open source solutions are integrated, including Hadoop, HDFS, Storm, and Oozie.

**(v) Service:**Data mining functions will be provided as service.

**(vi) Security/privacy/standard:**Security, privacy, and standard are very important to big data mining system. Security and privacy protect the data from unauthorized access and privacy disclosure. Big data mining system standard makes data integration, sharing, and mining more open to the third part of developer.

## 5. Conclusion:

The Internet of Things concept arises from the need to manage, automate, and explore all devices, instruments, and sensors in the world. In order to make wise decisions both for people and for the things in IoT, data mining technologies are integrated with IoT technologies for decision making support and system optimization. Data mining involves discovering novel, interesting, and potentially useful patterns from data and applying algorithms to the extraction of hidden information. In this paper, we survey the data mining in 3 different views: knowledge view, technique view, and application view. In knowledge view, we review classification, clustering, association analysis, time series analysis, and outlier analysis. In application view, we review the typical data mining application, including e-commerce, industry, health care, and public service. The technique view is discussed with knowledge view and application view. Nowadays, big data is a hot topic for data mining and IoT; we also discuss the new characteristics of big data and analyze the challenges in data extracting, data mining algorithms, and data mining system area. Based on the survey of the current research, a suggested big data mining system is proposed.

### Conflict of Interests…….

The authors declare that there is no conflict of interests regarding the publication of this paper.

## 6.References…….

[1] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, “Security of the internet of things: perspectives and challenges,” Wireless Networks, vol. 20, no. 8, pp. 2481–2501, 2014.

[2] C.-W. Tsai, C.-F. Lai, and A. V. Vasilakos, “Future internet of things: open issues and challenges,” Wireless Networks, vol. 20, no. 8, pp. 2201–2217, 2014.

[3] H. Jiawei and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2011.

[4] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, “A survey of multiobjective evolutionary algorithms for data mining: part I,” IEEE Transactions on Evolutionary Computation, vol. 18, no. 1, pp. 4–19, 2014.

[5] Y. Zhang, M. Chen, S. Mao, L. Hu, and V. Leung, “CAP: crowd activity prediction based on big data analysis,” IEEE Network, vol. 28, no. 4, pp. 52–57, 2014.

[6] M. Chen, S. Mao, and Y. Liu, “Big data: a survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.

[7] M. Chen, S. Mao, Y. Zhang, and V. Leung, Big Data: Related Technologies, Challenges and Future Prospects, SpringerBriefs in Computer Science, Springer, 2014.

[8] J. Wan, D. Zhang, Y. Sun, K. Lin, C. Zou, and H. Cai, “VCMIA: a novel architecture for integrating vehicular cyber-physical systems and mobile cloud computing,” Mobile Networks and Applications, vol. 19, no. 2, pp. 153–160, 2014.

[9] X. H. Rong, F. Chen, P. Deng, and S. L. Ma, “A large-scale device collaboration mechanism,” Journal of Computer Research and Development, vol. 48, no. 9, pp. 1589–1596, 2011.

[10] F. Chen, X.-H. Rong, P. Deng, and S.-L. Ma, “A survey of device collaboration technology and system software,” Acta Electronica Sinica, vol. 39, no. 2, pp. 440–447, 2011

IJSER

IJSER