Review of Computer Engineering Studies
Vol. 7, No. 2, June, 2020, pp. 42-47
Journal homepage: http://iieta.org/journals/rces

Data Summarization and Modelling

Raveendranathan C. Kalathil

Principal, Rajadhani Institute of Engineering and Technology, Rajadhani Hills, Nedumparambu P.O., Thiruvananthapuram 695 102, Kerala State, India

Corresponding Author Email: principal@riet.edu.in

https://doi.org/10.18280/rces.070204

Received: 25 March 2020
Accepted: 16 May 2020

Keywords:
data summarization, data analytics, big data, IoT, KDD, data science, data mining, machine learning, deep learning, cyber-physical systems

ABSTRACT

Data analytics find a myriad of applications in all walks of human endeavour. The modern world is characterized by huge amounts of data emerging from all human-made systems, including sensors, communication devices, signal processors and conditioners. The plethora of data originating from several sources often have to be cleaned and filtered to make it useful for further processing. In this paper, we discuss a very important domain in Big Data processing-namely, Data Summarization and its modelling. Big Data refers to the fairly huge chunks of data originating from several real-world sources, including those from all communication systems (including voice, text-data, images, and video), sensing devices (for example, those from sensors in a wired or wireless sensor network), digital data processors and several other natural sources including those from galactic sources from other planets and stars outside our galaxy. To effectively process the data, even with the almost seamless processing power of modern-day digital computers, one has to resort to Data Summarization. Modelling is another key component in visualizing the impact of such pre-processed data to arrive at meaningful conclusions on the processes which are being studied.

1. INTRODUCTION

The term "Data Science" is increasingly used as the term "Big Data". As Science implies systematic study, Data Science is precisely the scientific study, using statistical tools. Sometimes Data Science is categorized as the "systematic study of the organization and analysis, and properties of data and its applications in inference, including our confidence in the inference" [1]. However, Data Science is more than a mere analysis of data using statistical tools in several ways. Data is considered as the new oil. Arguably, the raw material we deal with in Data Science is largely heterogeneous and amorphous data. It can be text, images, and/or video. These raw data often originates from devices and systems having intricate relations among themselves. Thus, a specialist in Data Science needs to possess interdisciplinary skills set encompassing mathematics, statistics, databases, artificial intelligence, machine learning, and optimization. He/She should also possess a thorough knowledge of the art of framing problems to create successful solutions. Thus the essence of Data Science lies in obtaining insightful discoveries on "what patterns satisfy this data". An insight becomes actionable when its predictive power is high so that predictions become more and more accurate. Note that the emphasis on the quality of forecast is mainly strong in applications involving knowledge discovery in databases (KDD) and machine learning. If a learning paradigm is unpredictive, it is dealt with scepticism and is of little value. Data Summarization is one of the major data mining concepts. It interleaves techniques for obtaining a compact depiction of a huge set of data elements. Mathematical Statistics summarization methods of tabulating the mean and standard deviations are frequently employed in explicit data visualization and analysis, as well as programmed generation of reports. The need for more summarization of data in the mining process is because we are living in a digital era where data transfers occur at enormous speeds in reality and it is very much faster than the capability of a human being. The business community often work on a huge chunk of data which is derived from a myriad of sources including Social Networks and Media, newspapers, books, and storage devices positioned in the cloud. Very often summarization of such raw data poses challenges to the user. More often the user is not expecting the huge chunk of data. This is since when he/she retrieves data from their sources, the amount of data stored in the database becomes unpredictable.

Machine learning-based Data Summarization can be unsupervised, supervised, or semi-supervised and they all learn from data. In a supervised learning approach, there is a collection of raw data and their corresponding summaries created by humans. One can derive useful features of sentences can be learnt from the tagged data. A fairly huge amount of labelled or annotated data is needed for the learning process. Supervised learning algorithms include Support Vector Machine (SVM), Naïve Bayes classification, Mathematical Regression, Decision trees, Neural networks, and Multilayer Perceptrons [2-5]. In contrast, unsupervised systems do not require any training data. They generate the summary by accessing only the target documents. They try to discover the hidden structure in the unlabelled data. Thus, they are suitable for any newly observed data without any advanced modifications. Such systems apply heuristic rules to extract highly relevant sentences and generate a summary.

Clustering and Hidden Markov Model (HMM) are some of the examples of unsupervised learning techniques [6]. Genetic algorithms (GA) are also a type of machine learning approach. Genetic algorithm, being a search heuristic technique, works on the process of natural selection [7]. Note that semi-supervised learning techniques require both labelled and unlabeled data to generate an appropriate function or classifier.

2. THE ORIGINS OF DATA SUMMARIZATION AND ITS IMPLICATIONS

The essence of Data Mining lies in finding out specific patterns in huge raw data sets relating methods which are a mix of algorithms of statistics, database systems, and machine learning. Data Mining is an interdisciplinary subfield of computer science and statistics. It aims to extract meaningful information (using innovative mathematical and statistical tools) from a raw dataset and transform the information into an intelligible composition for future applications. Precisely, Data mining is the first analysis step of what is termed as Knowledge Discovery in Databases (KDD) processes. Recently, both of them, Data Mining and KDD are gaining lots of attention from researchers, industrialist and media persons. In most of the real-world systems, data is collected at a massive rate, aggregated and processed at a fairly colossal rate. Thus there an urgent need for computational tools to collect or extract meaningful information from the huge chunks of data. The underlying theories, techniques and tools are collectively termed as KDD.

The conventional course of action transforming raw data into processed information(or “knowledge”) rests on analysis by humans followed by elucidation [8]. Knowledge discovery in databases has a very long list of applications ranging from healthcare, to astronomy, to geology to computer communication. Several businesses use data to gain a competitive advantage against competitors, and to increase efficiency, thus giving more priceless services to users. It may be emphasized that the raw data we incarcerate about our milieu is the basic proof we use to framework postulates and models of the space we are part of. Because computers have enabled human beings to accumulate and process more data than we can digest ourselves, it is quite natural to turn to high-performance computing devices to enable us to develop significant patterns and components from the huge chunks of raw data. Hence, Knowledge Discovery in Databases is an attempt to address a problem faced by the modern digital information age which is the overload of raw data. As an example, one can site the SKICAT system used by astronomers to carry out image analysis, classification, and cataloguing of heavenly bodies from the images obtained from radio telescopes and other imaging devices [9]. The SKICAT system could surpass humans and conventional computational tools used for categorizing the fairly faint heavenly bodies.

Other applications of data mining and KDD include marketing, share-market investments, fraud detection, manufacturing, telecommunications and data cleaning. KDD was successfully used to monitor customer group patters to predict their purchase plans. Similarly, principles of data mining were successfully deployed by several companies to predict the stock market situations. Credit card frauds were successfully detected using data mining. Another example is the CASSIOPEE troubleshooting system, developed jointly by SNECMA and the General Electric Company (GEC), which was deployed by three major European airlines to detect and forecast issues in the Boeing 737 aircraft. Another example for KDD is the Telecommunications Alarm-Sequence Analyzer (TASA) developed jointly by three telephone networks [10] and a manufacturer of telecommunications equipment. The TASA system used an original intelligent algorithm for locating frequently occurring alarm episodes from the alarm flow and formulating them as rules. Yet another example is the MERGE-PURGE system developed for the identification and elimination of duplicate welfare claims at the state of Washington, USA [11], which was applied effectively on data from the welfare department. Finally, we can cite the case of the use case of intelligent agents employed to steer through an information-rich environment, which is an important type of data discovery systems. Though the idea of active triggers in the database domain has long been investigated by several researchers, truly successful implementations of this idea emerged only with the advent of the Internet [8, 9]. If one examines the history of the idea of figuring out meaningful patterns in raw data, one can see that the process has been named as data mining, data archaeology, knowledge extraction, data pattern processing, information discovery, and information harvesting. It is interesting to note that statisticians, Data Analysts, and the people from the Management Information Systems (MIS) community preferred the usage of the term Data Mining.

2.1 The processes involved in KDD

In simple terms, the KDD involves nine steps. The first step is formulating a knowledge base of the application domain and the pinpointing the purpose of the KDD processes from the angle of the user. The second step is building a target data set: opting a data set, or focusing on a compartment of data samples or variables, on which knowledge discovery is to be done. The third process involves clean-up and preprocessing of raw data, wherein the basic operations include removing noise and artefacts from the data if needed. This step precisely includes assimilating necessary information to model noise, deciding on strategies for handling missing data fields, and provisioning for known changes and time-sequence information.

The fourth step involves data reduction and projection, which means finding useful features to represent the data depending on the objectives of the task. The fifth step is dimensionality reduction or transformation. The sixth step is matching the goals of the KDD process (step 1) to a specific Data-Mining method. The seventh step is probing analysis and selection of proposition and model. It involves choosing the data mining algorithm(s) and selecting method(s) to be used for searching for data patterns.

The eighth one involves interpreting the mined patterns, possibly returning to any of the previous steps 1 through 7 for further iteration. The final, ninth step is acting on the discovered knowledge, either using the knowledge directly, or disseminating the knowledge into another system for further action, or simply documenting it and transmitting it to parties interested in it.

2.2 The importance of data summarization

The amount of data flow in real-world communication systems, including social media, is so huge that it is not humanly possible in real-time to summarize the data available in such situations. Especially, crowd-sourced data in text form from social media sites like Facebook, Instagram, Twitter and

Flickr are important sources of real-time information on ongoing events in today's world. This includes socio-political events, natural and manmade disasters, etc. On such sites, micro-blogs are usually posted so quickly and in such enormous volumes, that humans can't run through all the posts [12]. The huge chunk of data emanating from all such sources contribute to what is known as Big Data. Big Data Analytics has emerged as a science in itself. There are several situations wherein data summarization and modelling help to understand the meaning of information contained in Big Data.

The amount of data flow in real-world communication systems, including social media, is so huge that it is not humanly possible in real-time to summarize the data available in such situations. Especially, crowd-sourced textual data from social media sites like *Twitter* and *Flickr* are nowadays important sources of real-time information on ongoing events, including socio-political events, natural and manmade disasters, and so on. On such sites, *micro-blogs* are usually posted so rapidly and in such large volumes, that it is not feasible for human users to go through all the posts [12].

2.3 Challenges in data summarization

To effectively summarize data, one should know what kind of features an applied knowledge discovery system is expected to have. The following challenges are often faced by a data scientist while summarizing data:

- Managing different flavours of data; data variability.
- Effectiveness and scalability of algorithms used for data summarization.
- Certainty, usefulness, and clarity of results.
- Articulation of an assortment of data summarization requests and outcomes.
- Cooperative data mining and knowledge abstraction at multiple levels.
- Data mining from raw data originating at different input points.
- Safeguarding of data security and privacy.

It is quite interesting to note that some of these requirements are mutually conflicting in nature.

3. A STUDY OF EXISTING TOOLS AND TECHNIQUES USED FOR DATA ANALYTICS

The following tools are widely used for data summarization. Note that the list is not exhaustive; there are several other tools available. However, the below listed ten tools are somewhat very popular among the user community: Microsoft Excel, The R Programming language, Tableau Public, Python, SAS, Apache Spark, RapidMiner, KNIME, QlikView, and Splunk.

Each one of the above has its pros and cons. For example, Microsoft Excel is a simple data analytics tool, which has several built-in functions that can be used for data summarization. But, it can be used mainly for structured data. Microsoft Excel is a popular, rudimentary and extensively applied analytical tool by people from several industries, irrespective of the user being an expert in SAS, R or Tableau. The importance of Microsoft Excel lies in the fact that it is quite handy to use it whilst there is a necessity of data analytics on the in-house data of the user. Microsoft Excel can filter and summarize the data as per the needs of the client with a sample of pivot tables. Microsoft Excel also possesses more advanced Business Analytics features that aids in automatic detection of relationships, formulating measures for Data Analysis Expressions (DAX), and temporal grouping.

The R programming language is gaining lots of popularity in data summarization as a powerful tool. R is the principal data analytics tool very much popular in the industry. R programming language is extensively used for mathematical statistics and data modelling. One can easily manoeuvre the raw user data using R and present in incongruous ways. Compilers for the R programming language are available on a variety of Operating System (OS) platforms including Microsoft Windows, Linux, UNIX, and the Mac OS. The R programming language has a rich collection of 11,556 packages, which can be easily browsed by categories. R also comes with tools which help to install all packages automatically based on the needs of the user. R programming language can be integrated well with Big Data.

Tableau Public is a free software tool that can connect any corporate source of data including Microsoft Excel, Data Warehouse, or Web-based Data and can generate data visualizations such as a 2-D and 3-D maps, dashboards and so on. It can present with instantaneous updates on the web and can also be shared through social media or with the user. Tableau Public permits the user to download the file in diverse file formats. To visualize the true capabilities of Tableau Public, one should deploy it along with a very good data source. Arguably, the Big Data processing capabilities of Tableau Public is the best available in the market and it can make the analysis and visualization of data a simple task.

Another free, open-source, object-oriented scripting language tool which is easy to read, write, and sustain is the Python programming language. The credit for developing the Python language goes to Guido van Rossum in the late 1980s. Python can support both structured and functional programming methods. The users find it quite easy to learn and program in Python language as it is very akin to other structured languages such as PHP, Ruby, and JavaScript. The deep learning and machine learning libraries of Python language are augmented by Keras, Scikit-Learn, Python Deep Learning Library Theano, TensorFlow, and PlaidML. Yet another very important feature of Python programming language is that it is platform-independent. It can be assembled on platforms like the SQL Server, JSON, and MongoDB database. The Python programming language can easily use textual data too.

Another popular data manipulation language which is assuming a leading position Data Analytics is the SAS is a programming environment and language. The SAS programming language was developed by the SAS Institute in 1966. It was upgraded further in the 1980s and 1990s. The major merits of the SAS programming environment are easy accessibility and manageability. Data from any source can be analyzed using SAS. The SAS institute developed and showcased an enormously large set of products in 2011. This includes modules for customer intelligence and numerous SAS modules for social media, web, and marketing analytics. These products are extensively used for profiling users and their projections. Various SAS modules are capable of effectively forecast behaviours of customers, manage and optimize communications links among them.

The Apache Spark is another rapid data processing engine developed by the AMP Lab at University of California, Berkeley (UCB), in the year 2009. Apache Spark is an ultrafast extensive data processing engine and executes applications in Hadoop clusters. Apache Spark speeds up algorithms 100 times faster in solid-state memory and 10 times faster on the hard disks. Apache Spark programming environment is modelled on Data Science and its deployment seamlessly transform the applications effortlessly in Data Science. Apache Spark is very much accepted for implementing data pipelines and the development of models in machine learning. Apache Spark programming environment comes with a library termed as Machine Learning Library (MLlib). MLlib provides a highly advanced cluster of machine learning algorithms for recurring data science tools such as Collaborative Filtering, Clustering, Classification, and Regression.

Another very powerful integrated open-source Data Mining software suite, RapidMiner, was developed by the company of the same name in 2006. The company, RapidMiner is a global leader in offering services for predictive analysis and other advanced data analytics like text analytics, machine learning, data mining, and visual analytics without any programming [13]. RapidMiner integrated data science platform can seamlessly integrate with any data source types, including Microsoft Excel, Microsoft Access, MySQL, Microsoft SQL, Teradata, IBM SPSS, Dbase, Oracle, Sybase, IBM DB2, Ingres and so on. RapidMiner is extremely powerful and can generate Data Analytics based on real-life data. Also, with RapidMiner, the user can manipulate the formats and data sets for extrapolative analysis. The free version of RapidMiner Studio, which is restricted to 1 logical processor and 10,000 data rows, are offered under the Affero General Public License (AGPL). The server version of RapidMiner is termed RapidAnaltics.

A panel of software professionals from the University of Konstanz developed the Konstanz Information Miner (KNIME) in January 2004. The architecture of KNIME is designed with the three key principles: visual-interactive framework, modularity, and easy expandability (scalability). With the help of KNIME, the user is enabled to analyze and model the data through visual programming. KNIME is a principal open-source, reporting, and integrated Data Analytics tool. KNIME integrates various components for machine learning, deep learning and data mining through its modular data pipelines. The KNIME is available under the GNU General Public License (GNU GPL).

The first version of the Data Analytics software tool, QlikView, was released in 1994, by a Swedish company Qlik. QlikView's Associative Engine enables users to coalesce several data sources so that associations and connections can be created across the raw data. The unique features of QlikView include patented technology and in-memory data processing. The results are obtained very fast to the end-users. The QlikView provisions the data in the report itself. Using the QlikView, data associations are automatically done and the raw data can be compressed to almost 10% of its original size. In the output of QlikView, related data is indicated using a specific colour and unrelated data by another colour. Thus relationships among data sets are visually rendered.

Splunk was developed by Splunk Inc., in the USA in 2004. The *Splunk* is a software contrivance which is employed for analyzing, searching, monitoring, and visualizing the machine-generated data in real-time. It can read, analyze and monitor the different type of log files and stores data as events in indexers. Splunk was created as a search engine for the log files that are maintained in the infrastructure of a system. Splunk summarizes all textual log data and presents an effortless way to navigate from end to end. Thus a user can pull in all kinds of machine-generated log data, and perform all sorts of interesting statistical analysis on it, and present it in a myriad of formats. Typical use cases of the Splunk software suite is in analyzing the log data of a typical TCP/IP communication in a computer network and predict chances of outages in the network. With the data proliferation resulting from the deployment of IoT and Wireless Sensor Networks, the relevance of Splunk has been enhanced quite a lot.

4. DATA SUMMARIZATION AND MODELLING-A CONTEMPORARY PERSPECTIVE

Based on the specific requirements listed out by the end-user, a data model describes clearly how the data should be used. Note that information refers to processed data and Data Modelling enables one to determine the needs of the information. It is interesting to note that Data modelling differs according to the nature of the business. This is because the processes in business in each zone are different, and it is to be identified in the Data Modelling phase. The first phase is analyzing the situation, viz. assembling the data. Data Modelling practice is initiated by collecting the requirements. While the development of the Data Model is in progress, it is all the more important to communicate the requirements with the end-users and other stakeholders. The act of exploring the data-oriented structures is precisely Data Modelling. The applications of Data Modelling are very versatile. A most striking function of Data Modelling is to help the end-user understand the requirements of the information. Properly understanding the requirements of the information makes the dealings between the stakeholders ("developers and end-users") easier. In summary, Data Modelling enables end-users to define their requirements precisely. Then the system developers are capable to build up a system to meet up those specified requirements better.

The abstract representation of data structures needed for a database is called the Data Model. It is very powerful in rendering and communicating the requirements of the business. The Data Model is a visual representation of the nature of data, the rules of the business processes that are pertinent to the data, and ultimately how it will be structured in the database. The three major types of designs of the Data Model are conceptual (abstract) design, logical design and physical design. It may be noted that both the functional and the technical teams of a project use the Data Model. The community of analysts and the end-users in the business constitute the functional team. The other group of software developers and programmers form the technical team. It is the responsibility of the Data Modellers to formulate and design the Data Model to meet the demands of the functional team, at the same time satisfying the needs of the technical team. Data models are of four types:

- Conceptual (Abstract) Data Models – defines the topmost echelon associations between dissimilar entities.
- Business Data Models – Meets the distinctive needs of an explicit business. However, this is somewhat akin to abstract models.
- Logical Data Models – List out the precise attributes, entities, and associations implicated in a business process. This serves as the starting point for the design of the physical model.
- Physical Data Models – Characterize a database and application-specific realization of a logical model.

4.1 Multidimensional data modelling

A multidimensional data structure is often depicted as "a variation of the relational model that uses multidimensional structures to organize data and express the relationships between data". Jensen et al., state that "multidimensional data models visualize a fundamental data component for the given domain, which is a uniquely defined function of multidimensional values".

One of the fundamental and formidable design issues in a corporate information system is termed the "Tower of Babel Dilemma". The formation of a standard model for the whole business with diverse data interpretation of an organizational unit is called the "Newspeak solution". However, permitting numerous and mismatched data models to exist simultaneously may lead to the "Tower of Babel problem". To avoid conflicts, the system designers can either frame an enterprise-wide data model or construct multiple data models to meet the needs of each department. But then, due to miscommunication, issues can arise, and the information system may not work the way it is intended to work.

4.2 Agent-based models

In systems having multiple agents (or "multi-agent simulation or multi-agent systems"), a group of computational archetypes are employed that simulate the activities and interactions of autonomous agents, with an idea to assess their effects on the modelled system in totality. Another name for such an entity is Agent-Based Model (ABM). Agent-Based Model often intertwines principles of complex systems theory, game theory, computational sociology, emergence, multi-agent systems, and evolutionary programming. To introduce randomness, Monte Carlo Methods are employed. Another terminology used for ABM is "individual-based models". According to researcher Nigel Gilbert, "Agent-based Modeling is a new analytical method for social sciences which is quickly becoming popular". It may be noted that agent-based modelling is a mathematical method that uses computational tools that enable a researcher to construct, critically determine, and try out with models composed of agents that intermingle within a milieu. The following nine approaches on focus will help to model an agent-based system in general: accessibility, modularity, preciseness, complexity management, excitability, expressiveness, refinability, openness, and analyzability.

At this juncture, we can point out that data modelling is usually done in the perspective of an information systems project with pertinent methods and paraphernalia. In the representation and documentation of data, data modelling can be used very effectively. The data model can be employed as a route-map to visit the data points from the beginning to the end. Sometimes data modelling is said to be the modern Global Positioning System (GPS) for professionals in Information Technology (IT) business, making it easier for them to navigate through the ocean of data. One can make the process highly advanced by removing the redundant steps, thereby making the data very lean and eliminate all unnecessary steps in it.

5. FUTURE DIRECTIONS

With the advent of high-speed broadband Internet, which offers almost seamless connectivity to the World Wide Web (WWW), there is indeed a data overflow or overload on the Internet. In such a context, data summarization has gained hitherto nonexisting importance. The so-called information overload and data overload (data proliferation on the Internet) creates a huge demand for on-the-go (dynamic) text and data summarization. There is a need to summarize the data one comes across from newspaper articles, books, and periodicals, stories on the topic that appear in the media, reports on various events, scientific research papers, data, news, audio and video clips, resume plays, weather forecasts, stock market film and speech. As there is an immense potential for its exponential growth, several top universities the world over are working on data summarization. To cite a few examples, the National Centre for Text Mining (NaCTeM)-Manchester University, University of Missouri-St. Louis and Aarhus University-Denmark have been resolutely working on data summarization algorithms, tools and techniques.

Data summarization techniques can be classified into two major classes as abstractive and extractive summarizations. In extractive summarization, the most significant phrases and sentences are extracted from the text be summarized and a new group is formed to create a summary of the text, without any major changes in the original [14]. The sequential nature of the original textual document is most often maintained. In abstractive summarization, the meaning of the original text is understood. This is done by making use of linguistic methods to establish and examine the meaning. Thus, abstractive summarization produces a generalized summary. This summary provides the information in a precise, concise and compact way which usually requires the generation of extended language and information compression.

It may be concluded that abstractive summarization is a much more efficient form of summarization compared to extractive summarization. This is due to the reason that it compacts information from multiple documents to create a precise summary. It is highly popular in the sense that it can develop new sentences summarizing essential information from the original text. In general, an abstractive text summarizer renders the summarized information in a coherent form that is grammatically correct and easily comprehensible. Note that the quality of the text summary depends a lot on the readability or linguistic rigour of the summarized text.

6. CONCLUSION

It may be noted that with the advent of game-changing technologies like the Internet of Things (IoT) and Cyber-Physical Systems (CPS), the importance of data summarization and modelling has gained much more vigour. Several researchers are working across the globe to develop new tools and techniques for data summarization. In this paper, we have examined some of them. The proliferation of the applications of IoT, Cyber-Physical Systems, and the Semantic Web resulted in Big Data. The effective use of machine learning techniques enables the machine to poise questions which may arise from the human counterparts. Data summarization and modelling still have to traverse a long path even now. It is expected that the new paradigms in data sciences and computing will show-case several new developments in the years to come. The availability of novel deep learning and machine learning algorithms and high-performance computing platforms make the task of data summarization and modelling an easy process.

ACKNOWLEDGEMENT

The author wishes to place on record his immense thanks to his students and colleagues at Rajadhani Institute of Engineering and Technology, Rajadhani Hills, Thiruvananthapuram for their continued support. He would also acknowledge his indebtedness to his family and guide and guru Professor Mangalathu Hari Sankar.

REFERENCES

[1] Vasant, D. (2013). Data science and prediction. Communications of the ACM, 56(12): 64-73. https://doi.org/10.1145/2500499

[2] Cortes, C.,Vapnik, V. (1995). Support vector networks. Machine Learning, 20: 273-297. https://doi.org/10.1007/BF00994018

[3] Huang, Y.G., Li, L. (2011). Naïve Bayes classification algorithm based on small sample set. 011 IEEE International Conference on Cloud Computing and Intelligence Systems, Beijing, pp. 34-39. https://doi.org/10.1109/CCIS.2011.6045027

[4] Freedman, D.A. (2009). Statistical Models: Theory and Practice. Cambridge University Press.

[5] Fattah, M.A., Ren, F. (2009). GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Computer Speech and Language, 23(1): 126-144. https://doi.org/10.1016/j.csl.2008.04.002

[6] Lin, H.J., Yang, F.W., Kao, Y.T. (2005). An efficient GA-based clustering technique. Tamkang Journal of Science and Engineering, 8(2): 113-122. https://doi.org/10.6180/jase.2005.8.2.04

[7] Guo, H., Zhou, Y. (2009). An algorithm for mining association rules based on improved genetic algorithm and its application. 2009 Third International Conference on Genetic and Evolutionary Computing, Guilin, pp. 117-120. https://doi.org/10.1109/WGEC.2009.15

[8] Fayyad, U., Piatetsky-Shapiro, G., Smyth P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3): 37-54. https://doi.org/10.1609/aimag.v17i3.1230

[9] Fayyad, U., Djorgovski, S.G., Weir, N. (1996). From digitized images to on-line catalogs: Data mining a sky survey. AI Magazine, 17(2): 51-66.

[10] Mannila, H., Toivonen, H., Verkamo, A.I. (1997). Discovery of frequent episodes in event sequences. Data Mining and Knowledge Discovery, 1: 259-289. https://doi.org/10.1023/A:1009748302351

[11] Hernandez, M.A., Stolfo, S.J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery, 2: 9-37. https://doi.org/10.1023/A:1009761603038

[12] Dutta, S., Chandra, V., Mehra, K., Das, A.K., Chakraborty, T., Ghosh, S. (2018). Ensemble algorithms for microblog summarization. IEEE Intelligent Systems, 33(3): 4-14. https://doi.org/10.1109/MIS.2018.033001411

[13] Hofmann, M., Klinkenberg, R. (2014). RapidMiner: Data Mining Use Cases and Business Analytics Applications. CRC Press.

[14] Jin, H.D., Wong, M.L., Leung, K.S. (2005). Scalable model-based clustering for large databases based on data summarization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(11): 1710-1719. https://doi.org/10.1109/TPAMI.2005.226