International Journal of Advance Research in Science and Engineering
Vol. No.6, Special Issue (01), September 2017, BVCNSCS 2017
www.ijarse.com
IJARSE
ISSN 2319 - 8354

# A STUDY ON KNOWLEDGE EXTRACTION PROCESS AND ITS IMPACT WITH RESPECT TO COGNITIVE IOT

**Dr.kiran Kumar Dharanikota[1], Dr.K.Bharghavi[2]**

[1]*Prof. in CSE, Sphoorthy Engneering College, Hyderabad (India)*

[2]*Associate Prof in CSE , Sphoorthy Engneering College, Hyderabad (India)*

**ABSTRACT**

*Internet of Things (IOT) and knowledge discovery are the two sides of the coin and both go together. In the absence of one, there is no use of other. Paper focuses on data generating sources and Knowledge discovery from that data and tools which are useful for the discovery of the knowledge and technique, which are to be followed for the purpose of discovering meaningful data from the huge amount of data .It also focus on types of the data and data generative sources and impact of those tools and technique for knowledge discovery.*

*Internet of Things (IOT) word is very booming in technological market and everyone is talking about the term Smart city especially in India and with reference to keyword smart city, IOT comes with it. The Small word IOT but very big responsibility comes on the shoulders of the technical person to Play with it and extract the data from the IOT and it connecting things this paper is focusing on the tools and technique used for knowledge discovery.*

***Keywords: IOT, Knowledge Extraction, Tools, Data Sources, Smart city, Sensors.***

## I. INTRODUCTION

How to find a “thing” in the Internet of Things (IoT)?

The answer to this question will be the key challenge that IoT users and developers are facing nowadays and will face in the future. Current models of IoT are focused on initial vertical solutions which is very limited by hardware and software platforms and its support. With the sudden increase of IoT in the upcoming years as per predicted by Cisco, IBM and Gartner, there is a necessity for second thoughts about how IoT can deliver value to the end-user ?So currently its very much required to first discover the current IoT development to understand the “thing” and to discover the knowledge or meaningful data from the IoT and to provide the facility to the user to develop the IoT applications, services for bringing the application as a “smart thing” without any knowledge of the things we have to also focus on how discovery can make a major impact for the future of IoT and auxiliary, become a required component for IoT success story. So, we need Cognitive Internet of Things (CIoT). Rightly said that “we are data rich but information poor” as so services provided by the CIoT but it will not be Intelligent if there is not a proper knowledge discovery.

## II. COGNITIVE INTERNET OF THINGS

IoT will be a major source of big data driven by its velocity, variety, value and volume. The diverse IoT data will be in high demand by business and end--user applications and hence will have to be stored in widely distributed, heterogeneous.

In current era and research, we are going through the word Internet Of Things(IoT) which focuses on enabling too many general object to see, hear ,smell for own and make it link to share that observation but in current era, only connected is not the enough we need to go beyond that general object must provide facility to learn , think and to understand physical and social world by themselves and this the reason new paradigm comes in existences named as Cognitive Internet of Things(CIoT), to make powerful to IoT to develop brain for high level intelligence CIoT that has the capability to bridge the physical world (with objects, resources, etc.) and the social world (with human demand, social behavior, etc.), and enhance smart resource allocation, automatic network operation, and intelligent service provisioning.

## III. KNOWLEDGE DISCOVERY

- **Why Discovery?**

The diversity in things and the data produced by them pose significant challenges in satisfying the dream of a truly interconnected smart world of things. [8]

In the 21st century, the human beings are used in the different technologies to adequate in the society. Each and every day the human beings are using the vast data and these data are in the different fields .It may be in the form of documents, may be graphical formats ,may be the video ,may be records (varying array ) .As the data are available in the different formats so that the proper action to be taken. Not only to analyze these data but also take a good decision and maintain the data .As and when the customer will require, the data should be retrieved from the database and make the better decision .This technique is actually we called as a data mining or Knowledge Hub or simply KDD(Knowledge Discovery Process).The important reason that attracted a great deal of attention in information technology the discovery of useful information from large collections of data industry towards field of "Data mining" [5]

### 3.1 .Data Mining

"we are data rich but information poor". There is huge volume of data but we barely able to turn them into useful information and knowledge for managerial decision making in business. To generate information it requires massive collection of data. It is in different formats like audio/video, numbers, text, figures, Hypertext formats .for making the complete used of the data

Simple data retrieval or data extraction is not up to the mark. So, for that it required smart tool which will able to do the automatic data integration, summarization, extraction and pattern discovery in row data. With the vast amount of data stored in files, databases, and other repositories, it is very much important, to develop potential tool for analysis and interpretation of data and for the sake of pull out required knowledge that help in proper

Decision-making and so, only one solution to this is data mining. Data mining means retrieving the hidden information from huge database/dataset. It is a very powerful technology and likely to help organizations to spotlight on the most important information in their data warehouses [1,2,3,4]. Data mining tools forecast future trends and behaviors, it helps organizations to make proactive knowledge-driven decisions [2]. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by prospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining, popularly known as Knowledge Discovery in Databases (KDD), it is the non-trivial extraction of implicit, previously unknown and potentially useful information from data in databases [3, 5]. It is actually the process of finding the hidden information/pattern of the repositories. [1,3,5].

### 3.2. Knowledge Discovery Sources (Iot)

Data Mining and Knowledge Discovery in the Real World, A large degree of the current interest in KDD is the result of the media interest surrounding successful KDD applications, for example, the focus articles within the last two years in Business Week, Newsweek, Byte, PC Week, and other large-circulation periodicals. Unfortunately, it is not always easy to separate fact from media hype. Nonetheless, several well documented examples of successful systems can rightly be referred to as KDD applications and have been deployed in operational use on large-scale real-world problems in science and in business. In science, one of the primary application areas is astronomy. Here, a notable success was achieved by SKICAT, a system used by astronomers to perform image analysis, classification, and cataloging of sky objects from sky-survey images (Fayyad, Djorgovski, and Weir 1996). In its first application, the system was used to process the 3 tera bytes (1012 bytes) of image data resulting from the Second Palomar Observatory Sky Survey, where it is estimated that on the order of 109 sky objects are detectable. SKICAT can outperform humans and traditional computational techniques in classifying faint sky objects. See Fayyad, Haussler, and Stolorz (1996) for a survey of scientific applications. In business, main KDD application areas includes marketing, finance (especially investment), fraud detection, manufacturing, telecommunications, and Internet agents. Marketing: In marketing, the primary application is database marketing systems, which analyze customer databases to identify different customer groups and forecast their behavior. Business Week (Berry 1994) estimated that over half of all retailers are using or planning to use database marketing, and those who do use it have good results; for example, American Express reports a 10- to 15percent increase in credit-card use. Another notable marketing application is market-basket analysis (Agrawal et al. 1996) systems, which find patterns such as, “If customer bought X, he/she is also likely to buy Y and Z.” Such patterns are valuable to retailers. Investment: Numerous companies use data mining for investment, but most do not describe their systems. One exception is LBS Capital Management. Its system uses expert systems, neural nets, and genetic algorithms to manage portfolios totalling $600 million; since its start in 1993, the system has outperformed the broad stock market (Hall, Mani, and Barr 1996). Fraud detection: HNC Falcon and Nestor PRISM systems are used for monitoring credit card

fraud, watching over millions of accounts. The FAIS system (Senator et al. 1995), from the U.S. Treasury Financial Crimes Enforcement Network, is used to identify financial transactions that might indicate money

laundering activity. Manufacturing: The CASSIOPEE troubleshooting system, developed as part of a joint venture between General Electric and SNECMA, was applied by

three major European airlines to diagnose and predict problems for the Boeing 737. To derive families of faults, clustering methods are used. CASSIOPEE received the European first prize for innovative applications (Manago and Auriol 1996). Telecommunications: The telecommunications alarm-sequence analyzer (TASA) was built in cooperation with a manufacturer of telecommunications equipment and three telephone networks (Mannila, Toivonen, and Verkamo 1995). The system uses a novel framework for locating frequently occurring alarm episodes from the alarm stream and presenting them as rules. Large sets of discovered rules can be explored with flexible information-retrieval tools supporting interactivity and iteration. In this way, TASA offers pruning, grouping, and ordering tools to refine the results of a basic brute-force search for rules. Data cleaning: The MERGE-PURGE system was applied to the identification of duplicate welfare claims (Hernandez and Stolfo 1995). It was used successfully on data from the Welfare Department of the State of Washington. In other areas,

a well-publicized system is IBM's ADVANCED SCOUT, a specialized data-mining system that helps National Basketball Association (NBA) coaches organize and interpret data from NBA games (U.S. News 1995). ADVANCED SCOUT was used by several of the NBA teams in 1996, including the Seattle Supersonics, which reached the NBA finals. Finally, a novel and increasingly important type of discovery is one based on the use of intelligent agents to navigate through an information-rich environment. Although the idea of active triggers has long been analysed in the database field, really successful applications of this idea appeared only with the advent of the Internet. These systems ask the user to specify a profile of interest and search for related information among a wide variety of public-domain and proprietary sources. For example, FIREFLY is a personal music-recommendation agent: It asks a user his/her opinion of several music pieces and then suggests other music that the user might like (<http://www.ffly.com/>). CRAYON (http://crayon.net/>) allows users to create their own free newspaper (supported by ads); NEWSHOUND (<http://www.sjmercury.com/hound/>) from the San Jose Mercury News and FARCAST (<http://www.farcast.com/> automatically search information from a wide variety of sources, including newspapers and wire services, and e-mail relevant documents directly to the user. These are just a few of the numerous such systems that use KDD techniques to automatically produce useful information from large masses of raw data. See Piatetsky-Shapiro et al. (1996) for an overview of issues in developing industrial KDD applications.[6]

## IV.KNOWLEDGE DISCOVERY PROCESS

As we all are aware that discovering knowledge is not a simple thing. So, for that, following are the stages which are followed for Knowledge discovery.

The concept of a KDDM process model was originally discussed during the first workshop on KDD in 1989 (Piatesky-Shapiro, 1991). The main driving factor to define the model was acknowledgement of the fact that knowledge is the end product of a data-driven discovery process. One of the outcomes of the workshop was also

the acknowledgement of the need to develop interactive systems that would provide visual and perceptual tools for data analysis. Following this seminal event, the idea of a process model was iteratively developed by the KDD community over the several years that followed. Initial KD systems provided only a single DM technique, such as a decision tree or clustering algorithm, with a very weak support for the overall process framework (Zytow& Baker, 1991; Klosgen, 1992; Piatesky-Shapiro &Matheus, 1992; Ziarko et al., 1993; Simoudis et al., 1994). Such systems were intended for expert users who had understanding of DM techniques, the underlying data, and the knowledge sought. There was very little attention focused on the support for the layman data analyst, and thus the first KD systems had minimal commercial success (Brachman &Anand, 1996). The general research trends were concentrated on the development of new and improved DM algorithms rather than on the support for other KD activities.[7]

## V. KNOWLEDGE DISCOVERY TOOLS AND TYPES [11][12][14][15]

In this section we first provide a feature classification scheme to study knowledge discovery and data mining tools. We then apply this scheme to review existing tools that are currently available, either as a research prototype or as a commercial product. Although not exhaustive, we believe that the reviewed products are representative for the current status of technology.

A knowledge discovery and data mining tools require a tight integration with database systems or data warehouses for data selection, pre-processing, integrating, transformation, etc. Not all tools have the same database characteristics in terms of data model, database size, queries supported, etc. Different tools may perform different data mining tasks and employ different methods to achieve their goals. Some may require or support more interaction with the user than the other. Some may work on a stand-alone architecture while the other may work on a client/server architecture. To capture all these differences, we propose a feature classification scheme that can be used to study knowledge discovery and data mining tools. In this scheme, the tools' features are classified into three groups called general characteristics, database connectivity, and data mining characteristics which are described below. A survey of Knowledge Discovery and Data Mining process models[9][12]

### 5.1.1 Content Pre-processing

Content pre-processing is the process of converting text, image, scripts and other files into the forms that can be used by the usage mining. It's not hard to understand that the Web content can be used to filter the input to, or output from the pattern discovery algorithm [5]. R. Cooley also described how the page views play the important roles in the pre-processing. For the content of static page views, the pre-processing can be easily done by parsing the HTML and reformatting the information or running additional algorithm as desired. It would be much more complicated to the content of dynamic page views. To perform the pre-processing, the content of each page view must be "assembled", either by an HTTP request from a crawler, or a combination of template, script, and the database accesses.

### 5.1.2 Structure Pre-processing

The structure of a Web site is formed by the hyper links between page views. The structure pre-processing can be treated similar as the content pre-processing. However, each server session may have to construct a different site structure than others.

### 5.1.3 Usage Pre-processing

The inputs of the pre-processing phase may include the Web server logs, referral logs, registration files, index server logs, and optionally usage statistics from a previous analysis. The outputs are the user session file, transaction file, site topology, and page classifications. It's always necessary to adopt a data cleaning techniques to eliminate the impact of the irrelevant items to the analysis result. The usage pre-processing probably is the most difficult task in the Web Usage Mining processing due to the incompleteness of the available data [5]. Without sufficient data, it is very difficult to identify the users. The easiest way to improve the data quality is to get user cooperation, but it's not easy at all. There exists a conflict between the analysis needs of the analysts (who want more detailed usage data collected), and the privacy needs of the individual users (who want as little data collected as possible) [3]. However, the heuristics and statistics methods can be used to improve the quality of the Web usage data [11][12]. We may find some approach to solve the problem, but it is impossible to avoid the misidentification completely, since the Web is so dynamic and versatile. For example, any page view accessed through a client or proxy-level cache will not be "visible" from the server side, and the only verifiable method of tracking ached page views is to monitor usage from the client side [5].

The session identification is also a part of the usage pre-processing. The goal of it is to divide the page accesses of each user, who is likely to visit the Web site more than once, into individual sessions. The simplest way to do is to use a time-out to break a user's click-stream into session. The thirty minutes is used as a default time-out by many commercial products. Another problem is named as path completion, which indicates the determining if there are any important accesses missed in the access log. The methods used for the user identification can be used for path completion. The final procedure of the pre-processing is formatting, which is a preparation module to properly format the sessions or transactions. For the details of the data preparation for the Web mining, please refer to [4].

## 5.2 Pattern Discovery

This is the key component of the Web mining. Pattern discovery converges the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. According to the techniques adopted in this area, I will introduce this process in the separate subsections as follows.

### 5.2.1 Statistical Analysis

Statistical techniques are the most powerful tools in extracting knowledge about visitors to a Web site. The analysts may perform different kinds of descriptive statistical analyses based on different variables when analysing the session file. By analysing the statistical information contained in the periodic Web system report,

the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions [5].

**5.2.2 Association Rules**

In the Web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions [4]. The authors of [5] pointed that in the term of the Web usage mining, the association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for prefetching documents to reduce user perceived latency.

**5.2.3 Clustering**

Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies [4]. Clustering of users will help to discover the group of users, who have similar navigation pattern. It's very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to the individual users. The clustering of pages is useful for Internet search engines and Web service providers, since it can be used to discover the groups of pages having related content.

**5.2.4 Classification**

Classification is the technique to map a data item into one of several predefined classes. In the Web domain, Web master or marketer will have to use this technique if he/she want to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbour classifier, Support Vector Machines, etc. [5].

**5.2.5 Sequential Pattern**

This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another's in a time-ordered set of sessions or episodes. It's very meaningful for the Web marketer to predict the future trend, which help to place advertisements aimed at certain user groups. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis [5].

#### 5.2.6 Dependency Modelling

The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the Web domain. The modelling technique provides a theoretical framework for analysing the behaviour of users, and is potentially useful for predicting future Web resource consumption.

### 5.3 Pattern Analysis

Pattern Analysis is a final stage of the whole Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of Web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format that can be assimilate easily. This can be done with the help of some analysis methodologies and tools. There are two most common approaches for the pattern analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations [11][13]. All these methods assume the output of the previous phase has been structured. There are more techniques coming out in recent years, such as visualization etc.

This is also a fertilized research area. Although there are quite a few commercial analysis applications available and many more are free on the Web, most of them are dislike by users, considered too slow, inflexible, difficult to maintain and limited in the functionality. To develop the efficient, flexible, and powerful tools, lots of work need to be done for both researcher and developer. [11]

## VI. CONCLUSIONS

In this paper, we have discussed what knowledge discovery is and its categorization process. We have also focused on the needs and KD and its sources. Knowledge discovery or Text mining work not only limited with the structured data but now it's very much related to unstructured data (textual files). A lot of work can be done with reference to the CIoT and KDD but to further improve and extend, that work need to be done yet.

## REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R... "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5] International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2,

No.3, June 2012

[6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic SmythFrom Data Mining to Knowledge Discovery in Databases

[7] lukasz a. Kurgan1 and petr musilekThe Knowledge Engineering Review, Vol. 21:1, 1–24. 2006, Cambridge University Press doi:10.1017/S0269888906000737 Printed in the United Kingdom http://ubiquity.acm.org

[8] Article on investigation of knowledge discovery and data mining tools using a feature classification scheme

[9] Michael Goebel a survey of data mining and knowledge discovery software tools

[10] Web Mining and Knowledge Discovery of Usage Patterns CS 748T Project (Part I)Yan Wang February, 2000

[11] Neelamadhab Padhy1, Dr. Pragnyaban Mishra 2, and RasmitaPanigrahi3 International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.3, June 2012

[13] Cognitive Internet of Things: A New Paradigm beyond Connection Qihui Wu, Senior Member, IEEE, Guoru Ding, Student Member, IEEE, YuhuaXu, Student Member, IEEE, Shuo Feng, Zhiyong Du, Jinlong Wang, Senior Member, IEEE, and Keping Long, Senior Member, IEEE

[14] http://www.mariapinto.es/ciberabstracts/Articulos/Knowledge%20Discovery.htm

[15] http://www.hindawi.com/journals/ijdsn/2015/718390/tab1/

[16] http://www.hindawi.com/journals/ijdsn/2015/718390/