Analysis of Modern Trends and Approaches to Data Mining

Anoop Kumar Paharia

Sr. Lecturer, Govt. Womens Polytechnic College Gwalior MP.

Abstract: - The paper surveys extraordinary factors of statistics mining studies. Data mining is beneficial in obtaining knowledge from huge domain names of databases, statistics warehouses, and statistics marts. Different and modern regions of statistics mining also are discussed. Issues and demanding situations of statistics mining together with extraordinary open supply gear are addressed as well. Data mining is an essential and evolving studies place and utilized by biologists to statisticians and laptop scientists as well.

Keywords: - data mining, knowledge discovery in databases, areas and tools in data mining, challenges of data mining.

1. INTRODUCTION

Data mining is extracting records and understanding from the massive quantity of statistics. Data mining is an vital step in discovering understanding from databases. There are numbers of databases, statistics marts, statistics warehouses all over the world. If the statistics aren't analyzed to locate out the thrilling styles, then the statistics would turn out to be statistics tombs. Data miners search for the pearl withinside the sea of statistics. A statistics mining machine may generate masses of styles. Typically a small fraction of the styles are thrilling. Here the thrilling way useable, legitimate and novel. Moreover, it's miles nearly not possible to extract the thrilling hidden styles withinside the sea of statistics without the assist of statistics mining gear. There are seven steps in statistics mining. They are statistics cleaning, statistics integration, statistics selection, statistics transformation, statistics mining, understanding presentation and sample evolution[9]. Database era had developed from primitive file processing to the improvement of statistics mining gear and packages. The statistics can be collected from numerous packages along with technology and engineering, control, enterprise houses, authorities management and environmental control. Interesting statistics styles can be mined from spatial, time-related, text, biological, multimedia, net and legacy databases. Data mining facilitate control in choice making. The statistics mining task consists of the invention of idea descriptions, association, classification, prediction, clustering, fashion evaluation, deviation evaluation and similarity evaluation. Data mining in huge databases poses numerous necessities and demanding situations for the researchers and developers. A multidimensional statistics version is used for the layout of statistics warehouses and statistics marts. The middle of such version is statistics cube. Data cube includes huge set of records and quantity of dimensions. Dimensions are the entities on which an employer maintains records. By nature, they are hierarchical.

2. DIFFERENT AREAS OF DATA MINING

2.1 Web Mining

As there may be a massive quantity of statistics and records to be had withinside the World Wide Web, the statistics miners have a fertile region for internet mining. Web mining is statistics mining strategies for extraction of records from internet files and services. The contents of the internet are very dynamic. It is developing at a fast pace, and the records are constantly updated. Web mining can be divided into the subsequent subtasks [4].

1. Resource locating: locating files intended for the Web.
2. Information choice and preprocessing: Selection and preprocessing of the records retrieved from the Web.
3. Generalization: To find out the general styles from the man or woman in addition to multiple sites.
4. Analysis: Discovered styles are interpreted for significant knowledge. Web mining can be divided into Web Structure, Web Contents, and Web Access Patterns.

2.2 Text Mining

The time period textual content mining or KDT (Knowledge Discovery in Text) become first proposed with the aid of using Feldman and Dagan in 1996 [4]. The unstructured textual content may be mined the usage of facts retrieval, textual content categorization, or making use of NLP strategies as a preprocessing step. Text Mining includes many programs such that textual content categorization, clustering, locating styles and sequential styles in texts, computational linguistics, and affiliation discovery.

2.2 Spatial Data Mining

The spatial statistics mining offers with statistics associated to location. The explosion of geographically associated statistics for speedy improvement of IT, virtual mapping, faraway sensing, GIS needs for developing databases for spatial evaluation and modeling. Spatial statistics description, classification, association, clustering, trend, and outlier evaluation are the principle additives for spatial statistics mining.

2.3 Multimedia data mining

Multimedia information mining explores the interesting styles from databases associated with multimedia that manages a massive series of multimedia items. Multimedia items consist of audio, video, image, collection information, and hypertext information containing textual content, textual content markups, and linkages. Multimedia information studies focus on content-primarily based totally retrieval, similarity search, association, category, and prediction analysis.

2.4 Time series data mining

A time collection database adjustments its values and occasions with appreciation to time. Some of the examples of time-collection information are inventory marketplace information, business transaction information, dynamic manufacturing information, scientific remedy information, net web page get right of entry to a sequence and so on. The time collection studies include issues associated with similarity search, fashion analysis, mining sequential and periodic styles in time-related information.

2.5 Biological data mining

There is a massive garage of medical and organic records from DNA microarray records, genomic sequences, protein interactions in addition to sequences, digital fitness records, disease pathways, biomedical photos, and the listing are going on. In the medical context, biologists are seeking to find the organic methods which are the purpose of a disease. There are a few troubles associated with these high-dimensional organic records. These matters consist of noisy and incomplete records, integrating diverse reasserts of records and processing computer in-depth tasks. Biologists in addition to medical scientists used numerous records mining equipment to find out exciting and significant observations from a massive variety of heterogeneous records from one of a kind organic domains.

2.6 Educational data mining

Educational Data Mining (EDM) is an emerging studies location worried with the particular sorts of facts that come from instructional settings, and the use of the ones strategies to higher understand college students. Educational Data Mining focuses on growing new gear and algorithms for coming across facts styles. EDM develops strategies and applies strategies from statistics, machine mastering, and facts mining to research facts accumulated at some stage in coaching and mastering. New computer-supported interactive mastering strategies and gear have spread out possibilities to collect and examine scholarly facts, to find out styles and tendencies in the one's facts, and to make new discoveries and take a look at hypotheses approximately how college students learn. Data accumulated from on-line mastering structures can be aggregated over big numbers of college students and can comprise many variables that facts mining algorithms can probe for version building. Different scholar fashions are used for the prediction of destiny mastering conduct of the college students. Computational fashions are used primarily based totally on the scholar area and pedagogy.

2.7 Ubiquitous data mining (UDM)

The records miners have a brand new venture withinside the form of the ever-present get admission to via way of means of the usage of wearable computers, palmtops, mobileular phones, laptops. To extract hidden statistics from those gadgets calls for superior analysis. In the sector of UDM, communication, computation, security, etc. are a number of the factors. One of the targets of the UDM is to extract thrilling styles while minimizing the extra fee of the computing because of the above-noted factors. To put into effect records mining responsibilities like classification, clustering, associations, etc. are tough for ubiquitous gadgets. Small show areas, records control in the cell are a number of the demanding situations in this regards. The key troubles are the superior set of rules for cell and dispensed computing, records control troubles, records representation techniques, integration of those gadgets with database packages, UDM architecture, software program agents, agent interplay and packages of UDM [7].

2.7 Constraint-based data mining

Constraint-primarily based totally records mining is one in every of the growing regions wherein the records miners use the constraint for higher records mining. One of the packages of constraint-primarily based totally records mining are Online Analytical Mining Architecture (OALM) evolved by [8] and is designed for multidimensional in addition to constraint-primarily based totally mining primarily based totally on databases and records warehouses. Usually, records mining strategies lack person control. One shape of records mining is wherein the human involvement is there withinside the shape of constraints. There are numerous styles of constraints with their very own traits and purpose. They are know-how type, records, dimension/level, interestingness, rule constraints.

3. DATA MINING TOOLS

The following are the famous records mining open supply tools

3.1 Rapid Miner

This device is written in Java programming language, and it gives analytics of superior stage via its template-based framework. Users rarely must do any coding. Rapid Miner is able to manage diverse responsibilities like statistical modeling, predictive analytics, and visualization other than information mining responsibilities. Rapid- Miner gives studying schemes, fashions, and algorithms from WEKA and R scripts that make it extra powerful. This open supply is distributed beneath neath the AGPL open supply license and it is able to be downloaded from Source Forge. It is one in all the pleasant commercial enterprise analytics software. All the information mining responsibilities are bundled in a single unmarried suite [http://rapid-i.com/ content/view/181/190/].

3.2 WEKA

Weka becomes initially advanced in a non-Java model for reading agricultural information. Later, the Java model becomes advanced, and it has become an effective device for one-of-a-kinds information mining programs like predictive modeling and information analysis. This software program is unfastened below the GNU General Public License, that's a massive gain as in comparison to Rapid Miner. As it's far unfastened below the GNU General Public License that's a massive gain of it compared to its counterparts like Rapid Miner. It may be custom designed through the users. Most of the information mining jobs are supported through Weka. They are classification, clustering, regression, function extraction, visualization, etc. Its graphical person interface makes it a better-state-of-the-art devise for the information mining process.

3.3 R-Programming

Project R that is a GNU project, is written in C, FORTRAN and R Language. R language is used for writing plenty of modules of the software program itself. R programming software program is free, and it's also used for statistical computing and graphics. Data miners used R for growing statistical packages and studying the facts. In latest years the recognition of R had improved due to its ease of use and extensibility. R offers different statistical strategies that encompass linear and nonlinear modeling; facts mining procedures i.e. classification, clustering, time collection evaluation, and others. [http://www.r-project.org/][16].

3.4 Orange

Orange, a Python-based, effective, and open supply device for information mining customers for the purpose of information extraction. It has effective visual programming and Python scripting connected to it. It may be used for device gaining knowledge of as nicely as bioinformatics and textual content mining with the aid of using including add-ons. It’s filled with capabilities for information analytics. Orange has specialized accessories like Bioorange for bio-informatics [http://orange.biolab.si/features/].

3.5 KNIME

KNIME is able to act 3 principal tasks in records preprocessing. They are extraction, transformation, and loading. The records processing is carried out via way of means of permitting the meeting of nodes. It is an integration platform with sturdy records analytics and reporting. KNIME used modular records pipelining idea for system studying and records mining. It is used for commercial enterprise intelligence as well as monetary records mining. KNIME is easily extendible and maybe introduced a plug-in for specific jobs. This open supply is likewise written in Java and primarily based totally on Eclipse. The center model is composed of diverse records integration modules. Its research location now no longer simplest consists of pharmaceutical research however additionally commercial enterprise records, monetary intelligence, and CRM consumer records. [https://en.wikipedia.org/ wiki/KNIME].

3.6 NLTK

When it involves language processing tasks, NLTK is one of the principal players. NLTK is used for system learning, information mining, sentiment evaluation, and information scraping. It is likewise extensively used for language processing. Because it's written in Python, you possibly can construct packages on the pinnacle of it, customizing it for small tasks. NLTK performed a principal position as a coaching tool, look at tool, prototyping and may be used as a platform for superb research. [https://en.wikipedia.org/ wiki/Natural_Language_Toolkit].

4. LITERATURE REVIEW

There are plenty of statistics mining research round the globe

i). Students Mood recognition [5] was proposed by Christos N. Moridis et. al. for online Online self-evaluation test. Exponential common sense and formulation had been used in this regard. The inputs had been the student's preceding solutions and slide bar reputation. The exponential common sense variables had been a total wide variety of questions for the web self-evaluation test, student's goal, and slide bar value. Appropriate feedbacks are recorded based at the contemporary reputation of the moods of the students. Student's guide choice in their temper using slide bar with no automation is the limitation of the system.

ii). A novel weakly supervised cyber criminal network mining method [20] was proposed by Raymond Y.K. Lau et. al. The approach turned into primarily based totally on relationships each express and implicit amongst the cybercriminals. The messages published via way of means of these criminals at the social media had been the premise of this method. The set of rules used in this context turned into a context-touchy Gibbs sampling set of rules. The set of rules mined each transactional and collaborative semantics to discover the relationship amongst such criminals. The version used turned into a probabilistic generative version for extracting multi-phrase expressions. Two styles of cyber crook relationships had been set up in unlabeled messages. The technique used right here is the idea stage for the implicit semantics associated with the text.

iii). Shenghua Bao et. al. [18] proposed For discovering and connecting with social feelings primarily based totally on the on line files with feelings to assist the users to pick out associated files with the aid of using their emotional preferences. This is a trouble of document categorization. For such social affective textual content mining, a joint emotionsubject matter version was proposed with the aid of using introducing a further layer for such form of emotion modeling into Latent Dirichlet Allocation (LDA). Associate feelings with particular emotional context had been used instead of an unmarried term. The authors advanced an approximate inference version with the aid of using the use of Gibbs Sampling Algorithm. The version categorized textual content primarily based totally on extraordinary feelings along with touch, surprise, and empathy, etc. with the aid of using the use of social affective textual content as input.

iv). Luigi Lancieri et. al. [12] proposed A classification technique for Internet customers primarily based totally on their behavior on the internet to provide more desirable services. For this purpose, IP Address, timestamp, key phrases from the proxy cache, URL, categorized consumer behavior had been collected. Two specific forms of categorization algorithms had been used. One is called "difficult clustering" for partition and every other is "soft clustering" for locating overlapping clusters to organization customers. Hierarchical agglomerative clustering (HAC) became used for difficult clustering.

v). Li-Der Chou et. al. [11] proposed the usage of social media with the assist of cellular gadgets to create social community organization for the kids with developmental disabilities (CDD). Families with CDD, university, sanatorium, and basis came hand handy to proportion substantial information primarily based totally on on-line social community associated to childcare of such kids. The customers can access the utility with the assist of PDA, personal pc or cellular gadgets via way of means of putting in the utility on such gadgets.

vi). In [19], the authors used distributional functions of the textual content categorization that took into consideration the compactness and the placement of the first look at the phrase. Previous researchers had used 'bag of words' illustration and assigned a phrase with values and worried about whether or not the phrase regarded withinside the file or now no longer or the frequency of the phrase. The authors in their studies paintings explored different kinds of values which explicit distribution of phrases in a file. The distributional functions are utilized by a tf idf fashion equation and functions of various categories are mixed with the use of ensemble learning techniques. The authors proved experimentally that distributional functions are beneficial for textual content categorization. The categorization performance improves appreciably through the use of those functions with little extra value in assessment to traditional methods. The distribution function performances are more desirable in the case of lengthy files and while the writing fashion is casual.

vii). In [17], the authors designed net carrier advice systems. While designing net carrier advice systems, the focused studies hassle turned into to keep away from recommending unfair or bad offerings to the customers. The device must assist customers to pick out the proper carrier from the massive range of to be had net offerings. The widely advocated metric in this regard is the popularity of net offerings. The remarks ratings with the aid of using the customers are used for imparting carrier popularity score. Malicious and subjective personal remarks frequently result in bias that influences the popularity size of net offerings. In their studies work, they proposed a unique device for the same. Cumulative Sum Control Chart and Pearson Correlation Coefficient have been used to find malicious person remarks ratings. The device completed higher with the aid of using the use of Bloom filtering and proposed malicious remarks score prevention scheme. Extensive experiments have been conducted with the aid of using the use of 1.five million net carrier invocation records. The experimental outcomes confirmed that the fulfillment ratio of the net carrier recommendations can be stronger and the device may reduce the deviation of popularity size.

viii). In [13], the researchers proposed a novel a shrewd gadget which could be capable of detect the street injuries automatically, notify them by the usage of vehicular networks and estimate the the severity of the twist of fate primarily based totally on facts mining equipment and expertise interference. Various variables consisting of the automobile speed, the kind of automobiles involved, the effect speed, and the status of the airbag, etc. are used for measuring the the severity of the twist of fate. A prototype primarily based totally on off-the-shelf gadgets became advanced and validated it on the Applus + IDIADA Automotive Research Corporation facilities, displaying that this gadget can lessen the time had to alert and deploy emergency offerings significantly after a twist of fate take place. Three category algorithms had been used consisting of Decision Trees, Support Vector Machines, and Bayesian networks and had been as compared for satisfactory results. It became observed that the Bayesian version for the category is the satisfactory-appropriate version.

5. DATA MINING TECHNIQUES

Several information mining strategies are utilized in information mining tasks. Association, classification, clustering, prediction, sequential sample mining, etc. are information mining strategies.

5.1 Classification

The classification reveals policies that partition statistics into a few groups. The entrance for the class is the education set. The education set’s magnificence labels are already known. Classification assigns magnificence labels to unlabelled information primarily based totally on a version that acquires expertise from the education datasets. Such class is referred to as supervised getting to know because the magnificence labels are known. There are numerous class fashions. Some of the not unusual place class fashions are choice trees, neural networks, genetic algorithms, support vector machines, Bayesian classifiers. The software consists of credit score threat analysis, fraud detection, banking, and clinical software, etc. [4].

5.2 Clustering

Clustering is a technique of grouping statistics so that statistics inside the cluster have excessive similarity and multiple to statistics in different groups. Clustering algorithms can be used for organizing statistics, categorize statistics for version creation and statistics compression, outlier detection, etc. Many clustering algorithms had been evolved and are labeled as partitioning methods, hierarchical methods, density primarily based totally, and grid-primarily based totally methods. The datasets can be numerical or categorical. K-Means, hierarchical, DBSCAN, OPTICS, STING are a number of the famous statistics clustering algorithms [15].

5.3 Association Rule Mining

Association rule mining is a well-researched approach for coming across thrilling relations among variables in huge databases. In affiliation rule, the expression is of the form X=>Y, wherein X and Y are set of items. The a most important goal is to find out all of the guidelines that have guide and self-belief more than or same to minimal guide or self-belief in a database. Support way that how regularly X and Y occurs collectively as a percent of overall transactions. Confidence way that how a good deal a particular object is depending on another. There is no importance for the styles with low self-belief and guide. The customers can extract beneficial and thrilling facts from the styles with intermediate values of self-belief and guide. The affiliation rule mining algorithms include Apriori, AprioriTid, Apriori hybrid, and Tertius algorithms [15].

5.4 Neural Networks

Neural networks are a brand new computing paradigm this is stimulated through the organic apprehensive machine, consisting of the brain, to procedure information [15]. It includes growing mathematical systems with the cap potential to learn [4]. The Neural networks have the cap potential to extract significant and beneficial styles and traits from complicated statistics. It is relevant to real-international troubles specially withinside the case of industry. As the neural networks are excellent at figuring out styles or traits, they'll be relevant for prediction or forecasting needs. The machine consists of noticeably interconnected processing elements (neurons) running together to clear up a selected hassle. Artificial neural network (ANN) learns through example [13]. ANN is configured for a selected utility as classification, sample recognition, etc. via a mastering procedure. It will also be used for three-dimensional item recognition, hand-written phrase recognition, face recognition, etc. Neural networks have the downside of now no longer explaining the derived results. Another hassle is that it suffers from lengthy mastering times. As the statistics grow, the state of affairs turns into worse for that hassle.

5.6 Support Vector Machines

Support vector machines (SVM) belong to a new the magnificence of system getting to know algorithms and are primarily based totally on statistical getting to know theory [4]. The main the idea is to non-linearly map the records set into a excessive dimensional characteristic area and use a linear discriminator for the class of records. It is essentially used for regression, class and selection tree construction. SVMs pick the plane which maximizes the margin isolating the two classes. The margin is described as the gap among the isolating hyperplane to the closest the factor of A, plus the gap from the hyperplane to the closest factor in B, wherein A and B are two linearly separable sets. SVM has been used in many programs which include face detection, handwritten man or woman and digits recognition, speech recognition, picture, and information retrieval [14].

5.7 Genetic Algorithms

Genetic algorithms are a brand new paradigm in computing stimulated with the aid of using Darwin’s concept of evolution [4]. A populace of the person with a viable approach to trouble is created initially at random. Then the crossover is executed with the aid of using combining pairs of people to produce offspring of subsequent technology. A mutation procedure is used to alter the genetic shape of a few contributors of the latest technology randomly. The set of rules searches for an answer withinside the successive technology. When a superior answer is found or a few constant time is elapsed, the procedure comes to an end. Genetic algorithms are broadly used in troubles in which optimization is required.

6. REFERENCE:

i. Mo Hai, **"Survery of Clustering Algorithms for Big Data [J]"**, *Computer Science.*, vol. 43, no. 6A, pp. 380-383, 2016.

ii. S Jiang, J Ferreira and M C. Gonzalez, **"Activity-based human mobility patterns inferred from mobile phone data:a case study of Singapore[J]"**, *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 208-219, 2017.

iii. Adam Baba, Gouse Pasha, Shaik Althaf Ahammed, S. Nasira Tabassum, **“Introduction to Neural Networks Design Architecture”,** International Journal ofScientific & Engineering Research Volume 4, Issue 2, Februry 2013, ISSN 2229-5518.

iv. Arun K Pujari, Data Mining Techniques, University Press, 2013.

v. Christos N. Moridis and Anastasios A. Economides **“Mood Recognition during Online Self- Assessment Tests”** IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 2, NO. 1, JANUARY MARCH 2009.

vi. Manishaben Jaiswal,” **CLOUD COMPUTING AND INFRASTRUCTURE",** IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.4, Issue 2, Page No pp.742-746, June 2017,
DOI Member: 10.6084/m9.doi.one.IJRAR19D1251 Available at http://www.ijrar.org/viewfull.php?&p_id=IJRAR19D1251

vii. Eric Hsueh-Chan Lu, Wang-Chien Lee,Member, IEEE, and Vincent S. Tseng,Member, IEEE,” **A Framework for PersonalMobile Commerce Pattern Mining andPrediction”,** IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 24, NO. 5, MAY 2012

viii. H. Kargupta and A. Joshi, **“Data Mining to Go: Ubiquitous KDD for Mobile and Distributed Environments”,** KDD-2001, San Francisco, August 2001.

ix. J. Han, V.S. Lakshmanan and R T Ng,“**Constraint-based, Multidimensional Data Mining”,** COMPUTER (Special issue on DataMining), 32(8): 45-50, 1999

x. Jiawei Han and Micheline Kamber, **”Data Mining: Concepts and Techniques”**, MorganKaufmann Publishers, 2003.

xi. Kasun Wickramaratna, Student Member,IEEE ,Miroslav Kubat, Senior Member, IEEE,and Kamal Premaratne, Senior Member,IEEE,**” Predicting Missing Items in ShoppingCarts”,** IEEE TRANSACTIONS ONKNOWLEDGE AND DATA ENGINEERING,VOL. 21, NO. 7, JULY 2009.

xii. Li-Der Chou, Member, IEEE, Nien-Hwa Lai, Yen-Wen Chen, Member, IEEE, Yao-JenChang, Jyun-Yan Yang, Lien-Fu Huang,Wen-Ling Chiang, Hung-Yi Chiu, and Haw-Yun Shin **"Mobile Social Network Services for Families With Children With Developmental Disabilities"** IEEE TRANSACTIONSON INFORMATION TECHNOLOGY.

xiii. Manishaben Jaiswal **"GAME DEVELOPMENT PRINCIPLE, ARCHITECTURE AND METHODOLOGY",** International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.3, Issue 5, page no.267-270, May-2016,
DOI Member: 10.6084/m9.jetir.JETIR1912034
Available at: http://www.jetir.org/view?paper=JETIR1912034

xiv. Luigi Lancieri, Member, IEEE, and NicolasDurand **"Internet User Behavior: Compared Study of the Access Traces and Application to the Discovery of Communities"** IEEETRANSACTIONS ON SYSTEMS, MAN, ANDCYBERNETICS—PART A: SYSTEMS ANDHUMANS, VOL. 36, NO. 1, JANUARY 2006.

xv. Manuel Fogue, Piedad Garrido, Member, IEEE, Francisco J. Martinez, Member, IEEE,Juan-Carlos Cano, Carlos T. Calafate, and Pietro Manzoni, Member, IEEE,**" A Systemfor Automatic Notification and Severity Estimation of Automotive Accidents",** IEETRANSACTIONS ON MOBILECOMPUTING, VOL. 13, NO. 5, MAY 2014

xvi. Maya Nayak and Jnana Ranjan Tripathy**:"Pattern Classification Using Neuro Fuzzyand Support Vector Machine (SVM) – A Comparative Study",** International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5,May 2013.

xvii. H.H. Darji, B. Shah & M.K. Jaiswal. **"CONCEPTS OF DISTRIBUTED AND PARALLEL DATABASE",** International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN:2249-9555 vol. 2, no. 6, Dec. 2012, No;224, pg. no: 1150 available at https://www.academia.edu/42308519/CONCEPTS_OF_DISTRIBUTED_AND_PARALLEL_DATABASE

xviii. N. Mlambo, **"Data Mining: Techniques, Key Challenges and Approaches for Improvement"**,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.

xix. Paško Konjevoda and Nikola Štambuk**,"Open-Source Tools for Data Mining in Social Science,"** Theoretical and Methodological Approaches to Social Sciences and Knowledge Management, pp. 163-176 .

xx. Shangguang Wang, Member, IEEE, Zibin Zheng, Member, IEEE, Zhengping Wu, Member, IEEE, Fangchun Yang, Member,IEEE, Michael R. Lyu, Fellow, IEEE, **"Reputation Measurement and Malicious Feedback Rating Prevention in Web Service**

xxi. Recommendation Systems"**,** IEEE TRANSACTIONSON SERVICES COMPUTING, VOL., NO. , MARCH 2014.

xxii. Shenghua Bao, Shengliang Xu, Li Zhang,Rong Yan, Zhong Su, Dingyi Han, and Yong Yu **"Mining Social Emotions from Affective Text "**IEEE TRANSACTIONS ONKNOWLEDGE AND DATA ENGINEERING,VOL. 24, NO. 9, SEPTEMBER 2012.

xxiii. Xiao-Bing Xue and Zhi-Hua Zhou, SeniorMember, IEEE, **"Distributional Features for Text Categorization",** IEEE TRANSACTIONSON KNOWLEDGE AND DATAENGINEERING, VOL. 21, NO. 3, MARCH 2009.

xxiv. Y.K. Raymond, Lau SAR Yunqing Xia,Yunming Ye Shenzhen **"A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media"**, CHINA.