UGC JOURNAL NUMBER 44557

International Journal of Allied Practice, Research and Review
Website: www.ijaprr.com (ISSN 2350-1294)

Text Mining and its Applications

Dr. TariqHussain Sheikh[1] and Sujat Khan[2]
Lecturer in GDC Poonch, Jammu and Kashmir, India
Assistant Professor, GLDM Degree College Hiranagar, Jammu and Kashmir, India

Abstract-: Quick advance in computerized information procurement procedures have prompted tremendous volume of information. More than 80 percent of the present information is made out of unstructured or semi-organized information. The disclosure of suitable examples and patterns to examine the content records from enormous volume of information is a major issue. Content mining is a procedure of extricating intriguing and nontrivial designs from tremendous measure of content reports. There exist diverse systems and devices to mine the content and find profitable data for future expectation and basic leadership process. The determination of right and proper content mining system improves the speed and reductions the time furthermore, exertion required to extricate significant data. This paper quickly examine and break down the content mining strategies and their applications in different fields of life. Additionally, the issues in the field of content mining that influence the exactness and importance of comes about are recognized.

Keywords : ***Classification; Knowledge Discovery; Applications; Information Extraction; patterns***

I. INTRODUCTION

The span of information is expanding at exponential rates day by day. All kind of foundations, associations, and business enterprises are putting away their information electronically. A colossal measure of content is streaming over the web as advanced libraries, stores, and other printed data for example, websites, online networking system and messages [1]. It is testing errand to decide fitting examples and patterns to extricate significant information from this huge volume of information [2]. Conventional information mining apparatuses are unable to deal with printed information since it requires time and push to extricate data.

Content mining is a procedure to extricate intriguing and noteworthy examples to investigate learning from literary information sources [3]. Content mining is a multi-disciplinary field in view of data recovery, information mining, machine learning, insights, what's more, computational phonetics [3]. A few content mining systems like outline, characterization, grouping and so forth can be connected to extricate learning.

Content mining manages regular dialect content which is put away in semi-organized and unstructured organization [4]. Content mining systems are persistently connected in industry, the

scholarly community, web applications, web and different fields [5]. Application territories like web crawlers, client relationship administration framework, channel messages, item proposal investigation, extortion recognition, and web-based social networking investigation utilize content digging for sentiment mining, highlight extraction, notion, prescient, and slant investigation [6].

Gathering unstructured information, from various sources is accessible in various document arrangements, for example, plain content, website pages, pdf records and so forth. Pre-preparing and purging operations are performed to distinguish and expel peculiarities. Purifying procedure make a point to catch the genuine embodiment of content accessible what's more, is performed to expel stop words stemming (procedure of recognizing the base of certain word) and ordering the information [7]. Handling and controlling operations are connected to review and further clean the informational index via programmed preparing.

Example investigation is executed by Management Information Framework (MIS). Data prepared in the above advances are utilized to remove significant and important data for compelling what's more, convenient basic leadership and pattern investigation [8].

Extraction of significant data from a corpus of various records is a dreary and tedious undertaking. The choice of proper procedure for mining content diminishes the time and exertion to locate the important examples for investigation and basic leadership. The goal of this paper is to investigate diverse content mining strategies which help to perform content examination successfully and productively from substantial measure of information. Additionally, the issues that emerge amid content mining process are recognized.

II. THE REFLECTIVE PROCESS

Distinctive content mining methods are accessible that are connected for examining the content examples and their mining procedure [16]. Figure 3 demonstrates the Venn outline for the interrelationship among content mining procedures and their center usefulness. Archive characterization (content order, report institutionalization), data recovery (watchword seek /questioning and ordering), report grouping (state bunching), normal dialect handling (spelling remedy, lemmatization, syntactic parsing, and word sense disambiguation), data extraction (relationship extraction/interface examination), furthermore, web mining (web interface investigation) [6].

A. Information Extraction

Data Extraction (IE) is a procedure that concentrates significant data from expansive measure of content. Area specialists determine ascribes and connection as indicated by the space [17]. IE frameworks are utilized to separate particular properties what's more, substances from the archive and build up their relationship [18]. The removed corpus is put away into database for additionally handling. Accuracy and review process is utilized to check and assess the pertinence of results on the separated information. Inside and out and finish data about the pertinent field is required to perform data extraction procedure to accomplish more pertinent outcomes [19].

B. Information Retrieval

Data Retrieval (IR) is a procedure of extricating significant also, related examples as indicated by a given arrangement of words or expressions. There is a cozy relationship in content mining and data recovery for literary information. In IR frameworks, unique calculations are utilized to track the client's conduct and hunt significant information likewise [19]. Google and Yahoo look motors are

utilizing data recovery framework all the more every now and again to separate pertinent archives as per an expression on Web. These web indexes utilize inquiry based calculations to track the slants and achieve more huge outcomes. These web crawlers give client more pertinent and proper data that fulfill them as indicated by their requirements [8].

C. Natural Language Processing

Normal dialect handling (NLP) worries to the programmed preparing and examination of unstructured printed data. It perform diverse sorts of examination, for example, Named Element Recognition (NER) for contraction and their equivalent words extraction to discover the connections among them [10]. NER recognize every one of the occasions of determined question from a gathering of reports. These elements and their examples permit the distinguishing proof of relationship and other data to accomplish their key idea. Nonetheless, this strategy needs total word reference list for all named elements utilized for ID [9], [10]. Complex inquiry based calculations should be utilized to accomplish worthy comes about. In true, a solitary element has various terms like TV and Television. Some of the time, a gathering of progressive words have multi-word names to recognize the limits and resolve covering issues by utilizing order method. Ways to deal with manage NER ordinarily fall into four classes: vocabulary, run, factual based or blend of these drew closer.

NER frameworks have accomplished the pertinence level from 75 to 85 percent [20]. To remove equivalent word and contraction from printed information, co-referencing strategy is much of the time being used for NLP. Normal Dialects (NL) have parcel of complexities as a content extricated from various sources don't have indistinguishable words or shortening. There is a need to recognize such issues and make rules for their uniform distinguishing proof [21]. For instance, NER and co-referencing approaches set up a sensible relationship to remove and distinguish the part of individual in an association (utilize the name of a man on the double and after that utilization pronoun rather than name over and over) [22].

D. Clustering

Grouping is an unsupervised procedure to characterize the content records in bunches by applying diverse grouping calculations. In a bunch, comparative terms or examples are assembled extricated from different reports. Bunching is performed in top-down and base up way. In NLP, different sorts of mining apparatuses and systems are connected for the investigation on unstructured content. Distinctive systems of bunching are progressive, dispersion, thickness, centric, and k-mean [22].

E. Text Summarization

Content rundown is a procedure of gathering and delivering succinct portrayal of unique content archives [23]. Pre-preparing and handling operations are performed on the crude content for synopsis. Tokenization, stop word expulsion, furthermore, stemming techniques are connected for pre-preparing. Vocabulary records are produced at preparing phase of content rundown.

In past, programmed content outline was performed on the premise of event a specific word or expression in archive. Later on, extra strategies for content mining were presented with standard content mining procedure to enhance the significance furthermore, precision of results [11].

To outline the content reports - weighted heuristics strategy are separated by highlights by following particular standards. Sentence length, settled expression, passage, topical word, and capitalized word recognizable proof highlights can be executed and investigated for content summerization. Content synopsis methods can be connected on numerous records in the

meantime. Quality and kind of classifiers rely upon nature and subject of the content archives [24].

III. APPLICATION OF TEXT MINING

A. Digital Libraries

Various content mining strategies and instruments are being used to learn the examples and patterns from diaries and procedures from huge measure of vaults. These wellsprings of data help in the field of innovative work. Libraries are an incredible wellspring of data for the specialists and computerized libraries are trying to the centrality of their accumulation. It gives a novel strategy for arranging data in such a way that make it conceivable to accessible trillions of archives on the web. It gives a novel approach to compose data and make it conceivable to get to a great many reports on the web. Green-stone global computerized library that help different dialects and multilingual interfaces give a springy technique for separating reports that handle numerous organizations, i.e., Microsoft word, pdf, postscript, HTML, scripting dialects what's more, email messages [11]. It additionally bolsters the report extraction as varying media and picture organize along with content archives. In content mining process different operation are performed like archives determination, advancement, removing data and handling elements among the archives and creating natural co-referencing and synopsis [25]. Entryway, Net Owl and Aylien are as often as possible utilized instruments for content mining in advanced libraries.

B. Academic and Research Field

In instruction field, different content mining devices and systems are utilized to break down the instructive patterns in particular district, understudy's enthusiasm for particular field and business proportion [24]. Utilization of content mining in inquire about field help to discover and arrange look into papers and applicable material of various fields at one put. The utilization of k-implies bunching and different procedures help to distinguish the properties of pertinent data. Understudies execution in various subjects can be gotten to and how distinctive traits impact the choice of subjects [11], [26].

C. Life Science

Life science and social insurance ventures are creating vast measure of printed and numerical information with respect to patient’s record, sicknesses, pharmaceuticals, side effects and medications of ailments and some more. It is a major test to sift through a suitable what's more, important content to take a choice from a huge natural store [25]. The medicinal records contain fluctuating in nature, unpredictable, long and specialized vocabulary are utilized that make the information disclosure process exceptionally troublesome [27]. Content mining instruments in biomedical field gives a chance to separate significant data, their affiliation and surmising relationship among different illnesses, species, and qualities. Utilize of a fitting content mining apparatuses in medicinal field help to assess the viability of restorative medications that show viability by looking at changed illnesses, manifestations and their course of medicines [28]. Content mining use in biomarker revelation, pharmaceutical industry, clinical exchange examination, and preclinical safe poisonous quality investigations, patent focused insight what's more, arranging, mapping of qualities

ailments and investigating the directed distinguishing pieces of proof by utilizing different devices [20].

D. Social Media

Content mining programming bundles are accessible for dissecting web-based social networking applications to screen and examine the on the web plain content from web news, online journals, email and so forth. Content mining instruments help to recognize and dissect number of posts, likes and supporters on the online networking system. This sort of examination demonstrate the general population response on various posts, news and how it spread around. It demonstrates the conduct of individuals have a place to particular age gathering or groups having closeness and
variety in sees about a similar post [29], [30]

E. Business Intelligence

Content mining assumes a critical part in business insight that assistance associations and endeavors to dissect their clients furthermore, contenders to take better choices. It gives a more profound understanding about business and give data how to enhance the consumer loyalty and increase focused points of interest [31]. The content mining devices like IBM content investigation, Quick digger, and GATE help to take choices about the association that produce alarms about great and awful execution, showcase changeover that assistance to take therapeutic activities. It moreover helps in media transmission industry, business and trade applications and client chain administration framework [32].

IV. ISSUES IN TEXT MINING FIELD

Many issues happen amid the content mining procedure and impact the proficiency and adequacy of basic leadership. Complexities can emerge at the middle of the road phase of content mining. In preprocessing organize different principles and controls are characterized to institutionalize the content that makes content mining process productive.

Before applying design examination on the archive there is a need to change over unstructured information into middle of the road shape be that as it may, at this stage mining process has its own particular confusions. At some point genuine topic or information lose its significance due to the adjustment in the content arrangement [27]. Another major issue is a multilingual content refinement reliance that makes issues. Just couple of apparatuses are accessible that help numerous dialects [33]. Different calculations and strategies are utilized autonomously to help multilingual content. Since various critical records continue outside the content mining process since different apparatuses don't bolster them. These issues make a load of issues in information disclosure and basic leadership process. Taint genuine advantage is hard to achieve by utilizing the existing content mining strategies and apparatuses in light of the fact that its once in a while bolster multilingual reports [34]. Joining of space learning is an essential region as it performs particular operations on indicated corpus and attains desired results. In this circumstances area learning from which report corpus to be removed need to incorporate with the figuring capacities from which data must be accomplished. As per the necessities of the field, specialists are expected to work cooperatively from different spaces to extricate more viable, exact and precise outcomes [22], [27]. The utilization of equivalent words, polysems and antonyms in the archives make issues (obscurity) for the content mining instruments that take both in a similar setting. It is hard to classify the records when

accumulation of report is extensive what's more, created from different fields having a similar space. Shortened forms gives changed importance in various circumstances is likewise a major issue [35]. Shifting ideas of granularity change the setting of content as per the condition and area information. There is have to depict rules as per the field that will be utilized as a standard in the region and can be inserted in content mining instruments as a module. It involves bunches of exertion and time to create and send modules in all fields independently. To create modules top to bottom and legitimate information about the particular space will be required [34], [36]. Regular dialects have bunches of confusions in itself that make issue in content refinement strategies and the ID of element relationship. Words having same spelling yet give various importances, for instance, fly and fly. Content mining devices considered both as comparable while one is verb and other is thing. Syntactic principles as indicated by the nature and setting is still an open issue in the field of content mining [36].

V. CONCLUSION

The accessibility of gigantic volume of content based information require to be inspected to extricate significant data. Content mining methods are utilized to break down the intriguing and applicable data adequately and effectively from extensive measure of unstructured information. This paper introduces a short outline of content mining strategies that assistance to enhance the content mining process. Particular examples and groupings are connected all together to extricate helpful data by disposing of unessential subtle elements for prescient examination. Choice and utilization of right systems also, apparatuses as indicated by the area help to make the content mining process simple and effective. Area learning mix, differing ideas granularity, multilingual content refinement, and common dialect preparing equivocalness are real issues and challenges that emerge amid content mining process. In future inquire about work, we will center to outline calculations which will help to determine issues displayed in this work.

VI. REFERENCES

[1] R. Sagayam, A survey of text mining: Retrieval, extraction and indexingtechniques, International Journal of Computational EngineeringResearch, vol. 2, no. 5, 2012.

[2] N. Padhy, D. Mishra, R. Panigrahi et al., "The survey of data miningapplications and feature scope," arXiv preprint arXiv:1211.5723, 2012.

[3] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping the power of textmining," Communications of the ACM, vol. 49, no. 9, pp. 76–82, 2006.

[4] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text mining:predictive methods for analyzing unstructured information. SpringerScience and Business Media, 2010.

[5] S.-H. Liao, P.-H. Chu, and P.-Y. Hsiao, "Data mining techniques andapplications–a decade review from 2000 to 2011," Expert Systems withApplications, vol. 39, no. 12, pp. 11 303–11 311, 2012.

[6] W. He, "Examining students online interaction in a live video streamingenvironment using data mining and text mining," Computers in HumanBehavior, vol. 29, no. 1, pp. 90–102, 2013.

[7] G. King, P. Lam, and M. Roberts, "Computer-assisted keyword anddocument set discovery from unstructured text," Copy at http://j.mp/1qdVqhx Download Citation BibTex Tagged XML Download Paper,vol. 456, 2014.

[8] N. Zhong, Y. Li, and S.-T. Wu, "Effective pattern discovery for textmining," IEEE transactions on knowledge and data engineering, vol. 24,no. 1, pp. 30–44, 2012.

[9] A. Henriksson, H. Moen, M. Skeppstedt, V. Daudaravi˘cius, andM. Duneld, "Synonym extraction and abbreviation expansion withensembles of semantic spaces," Journal of biomedical semantics, vol. 5,no. 1, p. 1, 2014.

[10] B. Laxman and D. Sujatha, "Improved method for pattern discoveryin text mining," International Journal of Research in Engineering andTechnology, vol. 2, no. 1, pp. 2321–2328, 2013.

[11] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges,techniques and technologies: A survey on big data," Information Sciences,vol. 275, pp. 314–347, 2014.

[12] R. Rajendra and V. Saransh, "A Novel Modified Apriori Approachfor Web Document Clustering," International Journal of ComputerApplications, pp. 159–171, 2013.

[13] K. Sumathy and M. Chidambaram, "Text mining: Concepts, applications,tools and issues-an overview," International Journal of ComputerApplications, vol. 80, no. 4, 2013.

[14] P. J. Joby and J. Korra, "Accessing accurate documents by miningauxiliary document information," in Advances in Computing andCommunication Engineering (ICACCE), 2015 Second InternationalConference on. IEEE, 2015, pp. 634–638.

[15] Z. Wen, T. Yoshida, and X. Tang, "A study with multi-word featurewith text classification," in Proceedings of the 51st Annual Meeting ofthe ISSS-2007, Tokyo, Japan, vol. 51, 2007, p. 45.

[16] V. Gupta and G. S. Lehal, "A survey of text mining techniques andapplications," Journal of emerging technologies in web intelligence,vol. 1, no. 1, pp. 60–76, 2009.

[17] R. Agrawal and M. Batra, "A detailed study on text mining techniques,"International Journal of Soft Computing and Engineering (IJSCE) ISSN,pp. 2231–2307, 2013.

[18] D. S. Dang and P. H. Ahmad, "A review of text mining techniquesassociated with various application areas," International Journal ofScience and Research (IJSR), vol. 4, no. 2, pp. 2461–2466, 2015.[19] R. Steinberger, "A survey of methods to ease the development ofhighly multilingual text mining applications," Language Resources andEvaluation, vol. 46, no. 2, pp. 155–176, 2012.

[20] A. M. Cohen and W. R. Hersh, "A survey of current work in biomedicaltext mining," Briefings in bioinformatics, vol. 6, no. 1, pp. 57–71, 2005.

[21] E. A. Calvillo, A. Padilla, J. Mu¯noz, J. Ponce, and J. T. Fernandez,"Searching research papers using clustering and text mining," inElectronics, Communications and Computing (CONIELECOMP), 2013International Conference on. IEEE, 2013, pp. 78–81.

[22] B. L. Narayana and S. P. Kumar, "A new clustering technique on textin sentence for text mining," IJSEAT, vol. 3, no. 3, pp. 69–71, 2015.

[23] B. A. Mukhedkar, D. Sakhare, and R. Kumar, "Pragmatic analysis baseddocument summarization," International Journal of Computer Scienceand Information Security, vol. 14, no. 4, p. 145, 2016.

[24] R. Al-Hashemi, "Text summarization extraction system (tses) usingextracted keywords." Int. Arab J. e-Technol., vol. 1, no. 4, pp. 164–168, 2010.

[25] I. H. Witten, K. J. Don, M. Dewsnip, and V. Tablan, "Text mining in adigital library," International Journal on Digital Libraries, vol. 4, no. 1,pp. 56–59, 2004.

[26] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining modelfor higher education system," Europen Journal of Scientific Research,vol. 43, no. 1, pp. 24–29, 2010.

[27] A. Henriksson, J. Zhao, H. Dalianis, and H. Bostr¨om, "Ensembles ofrandomized trees using diverse distributed representations of clinicalevents," BMC Medical Informatics and Decision Making, vol. 16, no. 2,p. 69, 2016.

[28] I. Alonso and D. Contreras, "Evaluation of semantic similarity metricsapplied to the automatic retrieval of medical documents: An umlsapproach," Expert Systems with Applications, vol. 44, pp. 386–399,2016.

[29] C. Ding and H. Peng, "Minimum redundancy feature selection frommicroarray gene expression data," Journal of bioinformatics and computationalbiology, vol. 3, no. 02, pp. 185–205, 2005.

[30] Y. Zhao, "Analysing twitter data with text mining and social networkanalysis," in Proceedings of the 11th Australasian Data Mining andAnalytics Conference (AusDM 2013), 2013, p. 23.