The International Journal of Engineering and Science (IJES)
|| Volume || 9 || Issue || 10 || Series I || Pages || PP 52-56 || 2020 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

# Data mining analysis for text document protection: A Survey

Ch.VenkateswaraRao[1], Dr.S.indraneel[2], Dr.D.NatarajaSivan[3]
[1]*Research Scholar, Annamalai University, Chidambaram,Tamilnadu,India.*
*Assoc.Prof,Dept.of CSE,PITS,Tenali,A.P.*
[2]*Professor, St.Ann's Engg.college. Chirala,AP.*
*Asst.Prof.Dept.of.CSE, Annamalai University, Chidambaram, Tamilnadu, India.*

***----------ABSTRACT----------***
*In present days, keeping data securely is main priority for all industries, public organizations and other sectors. The data security is becoming main problem in rapid growing internet based technologies like Artificial intelligence, cloud computing and Internet of Things (IoT). The applications like smart city will store more data in cloud so, it is a big task to researchers to keep this data safe. This paper explains about different data mining techniques for text document data protection and analyzed these techniques based on type of data (medium), security, imperceptibility, capacity, robustness and their drawbacks. For easiness of understanding to the readers, this paper divided data mining approaches based on image, structural, linguistic and hybrid. All techniques analyzed in this paper provide security for data available in local as well as cloud.*
***Keywords-Data mining, security,IoT, smart applications,cloud data.***

Date of Submission: 09-10-2020 Date of Acceptance: 24-10-2020

## I. INTRODUCTION

These days, one of the real sorts of data on the planet is in computerized design. Both positive and negative parts of the advanced arrangement are advancing in the cutting edge computerized world. Positive viewpoints incorporate advancement in stargazing, restorative science, and innovations. Then again, relating to the negative viewpoints, the abuse of these advances raises numerous issues like copyright security and information control. Because of the cutting edge innovations, for example, rapid PC systems, and Internet, and so on various ways have been utilized to unlawful duplicate, redistribute and store the computerized substance effectively. It is important to verify advanced substance and ensures them against unapproved duplicate [1].

Information is moved on distributed computing as sound, video, picture and content. The possession confirmation and copyright insurance of information is a difficult errand. Information is the critical component in savvy urban areas which continues the foundation of information and causes individuals to access computerized substance. Computerized watermarking gives an answer for advanced substance copyright assurance and proprietorship check. A mystery message is put inside an advanced substance without bargaining significant information. This mystery data is utilized later for proprietorship ID. Advanced watermarking is classified into content watermarking, picture watermarking, sound watermarking and video watermarking. The greater part of the exploration has concentrated on picture, sound, and video. As of now, content watermarking has gotten notoriety because of enormous quantities of the content archive are delivered and shared [2]. The people, government officials, and military confronting information security issues which additionally influence the shrewd urban areas. Computerized distributers have rights however confronting numerous dangers, for example, illicit utilization of copyrights, information control, and redistribution of data [3].

Content archives are a piece of pretty much every association or organization, for example, reviews firms, banks, or any huge private or open company. These reports are as budget summaries, legitimate notes, birth endorsements, delicate degrees, grouped reports and assertions [4]. Be that as it may, the vast majority of the current methods produce twisting during watermark addition, which straightforwardly impacts on indistinctness. Besides, the majority of the current systems are not strong or absence of limit. Changing over a computerized record to another configuration has the danger of losing the implanted watermark. The test is to guarantee the innovation and copyright insurance of content record, which required an appropriate watermark method that is powerful against designing assaults, intangible, accomplish high implanting limit and verified. This issue can be tended to by another system which is proposed here to defeat the content watermarking flow difficulties.This is the Review paper on data mining.

## II. LITERATURE STUDY

Digital text document is present research era and it's started in 90's. This section describes various data mining approaches explained by various authors. These approaches are divided into four ways, they are image, structural, linguistic and hybrid.

### *A. Image based approach*

In [5], author explains different digital image water marking techniques, its requirements and applications. The main requirements for image water marking are capacity, robustness and transparency. Applications are copyright protection, content archiving, meta-data insertion, broadcast monitoring, Tamper Detection and Digital fingerprinting. Water marking techniques are Discrete Cosine Transformation (DCT)domain,Discrete Wavelet Transformation(DWT) domain, Discrete Fourier Transformation(DFT) domain,Fast Fourier Transformation (FFT) domain and Discrete Hadamard Transfoemation(DHT) domain approach. The main drawback in this paper is the author explains about only few water marking techniques and their capacity is low.

Manmeet Kaur and Kamna Mahajan [6] presented about review of various text image based water marking techniques. The conclusion of this paper is research done on text based water marking is limited and researchers have to focus on robustness and accuracy. Stefano Giovanni Rizzo et al [7] mainly focused on security of data based on homoglyph characters substitution for Latin symbols. Authors hava proposed a system for preserving data by embedding watermark with passwords for short texts. This is good approach in terms of preserving content but the paper does not concentrate on preserving social networks contents and other cloud data.

In [8] the authors have presented a crossover method dependent on zero watermarking. The watermark is changed over into a picture as a string and afterward implanted in the spread record. It has one disadvantage that huge stockpiling is required to store Certified Authority (CA) keys.

In [9] the authors have proposed a spatial area picture watermarking strategy for examined and printed records. The white and blue segments of the shading picture are utilized to implant the watermark. The presentation is examined on the bases of various filtering goals, printable materials, and quality, which demonstrates that the proposed strategy is indistinct.

### *B. Linguistic-based Approaches*

The linguistic-based approach consists of semantic and syntactic techniques which emphasize the semantic that is used for embedding the watermark and does not change the meaning of the text. Using a synonym substitution technique semantic approach is developed, which specifies that the words are exchanged with their synonyms for data hiding. In this technique, grammatical alternations are used for watermark embedding without affecting the original meaning of the text. The verb, adverb, noun, pronoun, adjective, preposition, acronyms, and conjunction are language parts which are used for the watermarking.

In [10] author proposed a technique which depends on consolidating highlights of Chinese content sentences. Each word sentence entropy is determined, and the heaviness of each sentence is additionally acquiring through entropy. The proposed technique performs well in term of designing assaults. In [11] author proposed a phonetic methodology for content watermarking which dependent on attributes of exposition works. Agent words are utilized to create a watchword, center action word set and corresponding component of modifiers. Action words, descriptive word, thing, and qualifier are utilized for inserting watermark. The proposed system has a low implanting limit.

### *C. Structural based approach*

In these systems, basic bits are coordinated into the structure or normal for the content. The spaces among lines and words are used for watermark implanting. This methodology doesn't tackle the issue of possession validation and for a wide range of content archives. On the off chance that the spaces between words, lines, and sections are evacuated then shrouded information will be demolished. The investigation of line move coding is the first gathering of three lines, the center line is 1/300 inches moved down. In Syntactic-based methodology, this system likewise held all-regular printed highlights upheld by the advances of Natural-language programming (NLP) strategies and assets.

M. Yingjie et al [12] Propose a procedure for the Arabic language which uses the Kashida augmentation character and additional little whitespace for watermarking. The proposed procedure isn't strong against designing assaults, in such a case that the spaces between the words are expelled, the concealed data will be lost.

A. Taha et al [13] acquainted another methodology with zero content watermarking, which depends on the probabilistic model. The separating among words and line are utilized for watermarking. When contrasted with different methodologies, the proposed strategy performs better in reordering assaults and vigor. F. M. Ba-

Alwi et al [14] proposed a novel strategy, where first the watermarking data is scrambled through the Caesar figure with the client key at that point makes the gatherings of the message and pressing into plain message. Through examinations, the proposed plan isn't intangible.

In [15] author utilizes a method dependent on whitespaces between the words, which isn't vigorous against designing assaults and low inserting limit. The primary disservice of this is huge quantities of spaces are required to shroud the mystery message. In [16] author proposed a strategy which intends to ensure the information moved between the coherent, physical and IoT framework virtual segments, which consolidates the utilization of current advances of data security in the plan and activity of IoT frameworks. B. Usmonov et al [17] exhibited a design for secure online wellbeing applications utilizing IoT, Big Data and cloud combination to empower remote checking. Cloud View Ex-lead approach is utilized as a pursuit stage which offers access to framework level data for both on the web and undertaking based hunt applications. In [18] author propose an auxiliary methodology dependent on Font-Code, where as opposed to changing content letters the glyphs of the text styles are utilized for implanting watermark. The proposed calculation is strong and intangible however has Low limit and relevant for one text style family.

*D. Hybrid approaches*

A crossover approach has been created to join various ways to deal with content watermarking. These methods are viewed as vigorous and pertinent to wide content reports [19]. In [20] author Presented a technique for Arabic content dependent on pseudo-space. The pseudo-space disengages associated letters are utilized for watermarking. The proposed technique is intangible and vigorous against designing and altering assaults however can't hold against retyping assault. In [21] author proposed another system to shroud data that covers instant messages in the content utilizing the Omega system structure. . M. Hamdan and A. Hamarsheh [22], recommended the delicate watermark plan to ensure the honesty of the information in the IoT.

## III. ANALYSIS OF WORK

**Table 1:** Comparative analysis on various papers.

| Ref no | Medium | Security | Robustness | Capacity | Imperceptibility | Drawbacks |
|---|---|---|---|---|---|---|
| [11] | Text | High | High | Low | High | Low capacity |
| [12] | Text | High | High | Low | Medium | Low capacity |
| [13] | Text | Medium | Low | High | Medium | Low Robust against attacks |
| [15] | Text | High | Medium | High | Low | Low Robust against attacks |
| [16] | Text | Medium | Low | Low | High | Low Robust against attacks |
| [19] | Text | High | High | Low | Medium | Low capacity |
| [21] | Text | Medium | Medium | High | High | Low Robust against attacks |
| [22] | Text | High | High | Low | Medium | Low capacity |
| [24] | Text | Medium | Low | High | Low | Low Robust against attacks |
| [25] | Text | --- | Low | High | High | Low Robust against attacks |
| [26] | Text | High | Medium | Low | High | Low Robust against attacks |
| [27] | Text | Medium | Medium | Low | High | Capacity and robustness problem |
| [28] | Text | Medium | High | Medium | High | Low capacity |
| [29] | Image and text | High | High | Low | Low | Low Imperceptibility |
| [30] | Text | Low | Low | High | High | Not Robust |
| [31] | Text | Medium | Medium | High | Low | Low Robust against attacks |
| [32] | Text | High | Medium | Medium | High | Low Robust against attacks |
| [33] | Text | Medium | High | Low | High | Low capacity |
| [34] | Text | Medium | Low | High | Low | Low Imperceptibility |
| [35] | Text | High | High | Low | Medium | Low capacity |
| [36] | Text | Medium | High | High | Low | Low Imperceptibility |
| [37] | Text | High | Medium | High | Low | Low Imperceptibility |

This section explains about analysis of various papers by comparing type of data (medium), security, imperceptibility, capacity, robustness and their drawbacks. Table 1 explains al l these parameters. From the analysis it is clear that for text data security should be high for local as well as cloud, capacity must be more, high robustness and high imperceptibility. The main drawbacks in the analyzed papers are their capacity and robustness.

## IV. CONCLUSION

This paper explained about different data mining techniques used for text data productione analyzed these techniques based on type of data (medium), security, imperceptibility, capacity, robustness and their drawbacks. For easiness of understanding to the readers, this paper explained data mining approaches based on image, structural, linguistic and hybrid. For the analysis we considered and explained different papers only on text type of data. This text data capacity, robustness, security and imperceptibility must be high and protection of data must be for local system and cloud system.

## REFERENCE


[1]. M. Zeeshan, S. Ullah, S. Anayat, R. G. Hussain, and N. Nasir, ``A review study on unique way of information hiding: Steganography," *Int. J. Data Sci. Technol.*, vol. 3, no. 5, p. 45, 2017.

[2]. A. S. Panah, R. Van Schyndel, T. Sellis, and E. Bertino, ``On the properties of non-media digital watermarking: A review of state of the art techniques," *IEEE Access*, vol. 4, pp. 2670_2704, 2016.

[3]. M. Pal, ``A survey on digital watermarking and its application," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 1, pp. 153_156, 2016.

[4]. U. Khadim, A. Khan, B. Ahmad, and A. Khan, ``Information hiding in text to improve performance for word document," *Int. J. Technol. Res.*, vol. 3, no. 3, p. 50, 2015.

[5]. Y. Hao, Q. F. L. Chuang, and D. Rong, ``A survey of digital watermarking," *J. Comput. Res. Develop.*, vol. 7, pp. 1093_1099, 2005.

[6]. M. Kaur and K. Mahajan, ``An existential review on text watermarking techniques," *Int. J. Comput. Appl.*, vol. 120, no. 18, pp. 1_4, 2015.

[7]. S. G. Rizzo, F. Bertini, and D. Montesi, ``Content-preserving text watermarking through unicode homoglyph substitution," in *Proc. 20th Int. Database Eng. Appl. Symp.*, 2016, pp. 97_104.

[8]. O. Tayan, M. N. Kabir, and Y. M. Alginahi, ``A hybrid digital-signature and zero-watermarking approach for authentication and protection of sensitive electronic documents," *Sci. World J.*, vol. 2014, Aug. 2014, Art. no. 514652.

[9]. K. Thongkor and T. Amornraksa, ``Digital image watermarking for printed and scanned documents," *Proc. SPIE*, vol. 10420, Jul. 2017,Art. no. 104203O.

[10]. M. T. Ahvanooey *et al.*, ``A comparative analysis of information hiding techniques for copyright protection of text documents," *Secur. Commun. Netw.*, vol. 2018, pp. 1_22, 2018.

[11]. Y. Liu, Y. Zhu, and G. Xin, ``A zero-watermarking algorithm based on merging features of sentences for Chinese text," *J. Chin. Inst. Eng.*, vol. 38, no. 3, pp. 391_398, Apr. 2015.

[12]. M. Yingjie, L. Huiran, S. Tong, and T. Xiaoyu, ``A zero-watermarking scheme for prose writings," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discovery (CyberC)*, Oct. 2017, pp. 276_282.

[13]. A. Taha, A. S. Hammad, and M. M. Selim, ``A high capacity algorithm for information hiding in Arabic text," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published.

[14]. F. M. Ba-Alwi, M. M. Ghilan, and F. N. Al-Wesabi, ``Content authentication of english text via Internet using zero watermarking technique and Markov model," *Int. J. Appl. Inf. Syst.*, vol. 7, no. 1, pp. 25_36, 2014.

[15]. Y. Zhang, H. Qin, and T.Kong, ``Anovel robust textwatermarking forword document," in *Proc. 3rd Int. Congr. Image Signal Process. (CISP)*, vol. 1, Oct. 2010, pp. 38_42.

[16]. O. W. Liang and V. Iranmanesh, ``Information hiding using whitespace technique in Microsoft word," in *Proc. 22nd Int. Conf. Virtual Syst. Mul- timedia (VSMM)*, Oct. 2016, pp. 1_5.

[17]. B. Usmonov, O. Evsutin, A. Iskhakov, A. Shelupanov, A. Iskhakova, and R. Meshcheryakov, ``The cybersecurity in development of IoT embedded technologies," in *Proc. Int. Conf. Inf. Sci. Commun. Technol. (ICISCT)*, 2017, pp. 1_4.

[18]. G. Suciu *et al.*, ``Big data, Internet of Things and cloud convergence_An architecture for secure E-health applications," *J. Med. Syst.*, vol. 39, no. 11, p. 141, 2015.

[19]. C. Xiao, C. Zhang, and C. Zheng, ``FontCode: Embedding information in text documents using glyph perturbation," *ACM Trans. Graph.*, vol. 37, no. 2, p. 15, 2018.

[20]. Z. Jalil, ``Copyright protection of plain text using digital watermarking," FAST Nat. Univ. Comput. Emerg. Sci., Islamabad, Pakistan, Tech. Rep. 1059, 2010.

[21]. R. A. Alotaibi and L. A. Elrefaei, ``Improved capacity Arabic text watermarking methods based on open word space," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 30, no. 2, pp. 236_248, 2018.

[22]. A. M. Hamdan and A. Hamarsheh, ``AH4S: An algorithm of text in textsteganography using the structure of omega network," *Secur. Commun. Netw.*, vol. 9, no. 18, pp. 6004_6016, 2017.

[23]. G. Zhang, L. Kou, L. Zhang, C. Liu, Q. Da, and J. Sun, ``A new digital watermarking method for data integrity protection in the perception layer of IoT," *Secur. Commun. Netw.*, vol. 2017, Oct. 2017, Art. no. 3126010.

[24]. A. A.-A. Gutub, F. Al-Haidari, K. M. Al-Kahsah, and J. Hamodi, ``E-Text watermarking: Utilizing `Kashida' extensions in arabic language electronic writing," *J. Emerg. Technol. Web Intell.*, vol. 2, no. 1, pp. 48_55, 2010.

[25]. Y.-W. Kim, K.-A. Moon, and I.-S. Oh, ``A text watermarking algorithm based on word classi_cation and inter-word space statistics," in *Proc. ICDAR*, 2003, pp. 775_779.

[26]. H. Yang and A. C. Kot, ``Text document authentication by integrating inter character and word spaces watermarking," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, vol. 2, Jun. 2004, pp. 955_958.

[27]. Y. M. Alginahi, M. N. Kabir, and O. Tayan, ``An enhanced Kashida-based watermarking approach for Arabic text-documents," in *Proc. Int. Conf. Electron., Comput. Comput. (ICECCO)*, Nov. 2013, pp. 301_304.

[28]. Y. Meng, T. Guo, Z. Guo, and L. Gao, ``Chinese text zero-watermark based on sentence's entropy," in *Proc. Int. Conf. Multimedia Technol. (ICMT)*, Oct. 2010, pp. 1_4.

[29]. Z. Jalil and A. M. Mirza, ``Text watermarking using combined image-plustext watermark," in *Proc. 2nd Int. Workshop Educ. Technol. Comput. Sci. (ETCS)*, vol. 1, Mar. 2010, pp. 11_14.
[30]. R. J. Jaiswal and N. N. Patil, ``Implementation of a new technique for web document protection using unicode," in *Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES)*, Feb. 2013, pp. 69_72.
[31]. W. Cheng, H. Feng, and C. Yang, ``A robust text digital watermarking algorithm based on fragments regrouping strategy," in *Proc. IEEE Int. Conf. Inf. Theory Inf. Secur. (ICITIS)*, Dec. 2010, pp. 600_603.
[32]. N. Mir, ``Copyright for web content using invisible text watermarking," *Comput. Hum. Behav.*, vol. 30, pp. 648_653, Jan. 2014.
[33]. Y. Meng, L. Gao, X. Wang, and T. Guo, ``Chinese text zero-watermark based on space model," in *Proc. 3rd Int.Workshop Intell. Syst. Appl. (ISA)*,May 2011, pp. 1_5.
[34]. Y. M. Alginahi, M. N. Kabir, and O. Tayan, ``An enhanced Kashida-based watermarking approach for increased protection in Arabic text-documents based on frequency recurrence of characters," *Int. J. Comput. Elect. Eng.*, vol. 6, no. 5, p. 381, 2014.
[35]. Q. Wen, Y. Wang, and P. Li, ``Two Zero-Watermark methods for XML documents," *J. Real-Time Image Process.*, vol. 14, no. 1, pp. 183_192, 2018.
[36]. M. Kuribayashi, T. Fukushima, and N. Funabiki, ``Data hiding for text document in PDF _le," in *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, vol. 81. Cham, Switzerland: Springer, 2018, pp. 390_398.
[37]. L. Tan, K. Hu, X. Zhou, R. Chen, and W. Jiang, ``Print-scan invariant text image watermarking for hardcopy document authentication," *Multimedia Tools Appl.*, pp. 1_23, 2018.

Ch.VenkateswaraRao, et. al. "Data mining analysis for text document protection: A Survey." *The International Journal of Engineering and Science (IJES)*, 9(10), (2020): pp. 39-51.