A Method for finding threatened web sites through crime data mining and sentiment analysis

K.V. Daya Sagar¹, Ch Shyam Krishna², G. Lalith Kumar³, P. Surya Teja⁴, G. Charless Babu

¹Associate Professor, Department of electronics and computers, K.L.E.F, Guntur.
²Department of electronics and computers, K.L.E.F, Vijayawada, India
³Department of electronics and computers, K.L.E.F, Vijayawada, India
⁴Department of electronics and computers, K.L.E.F, Vijayawada, India

Malla Reddy Engineering College
*Corresponding author E-mail: sugar.tadepalli@gmail.com

Abstract
A Fast growth on the internet provided an opportunity for crime to develop in it. This includes using technology for planning illegal activities and message passing during the crime. It also includes replacing the documents with others which results in different effects. Majority of this done in the form of text. There may be a chance of using regular websites for doing crimes. This paper helps to find such websites through crime data mining and sentiment analysis. Sentiment analysis helps to find the regularly used sites of a user. Crime data mining helps to find the illegal text on the website. By combining both these techniques we can find the most threatened and most regularly used websites for the crime.

Keywords: Cybercrime, Data Mining, Semantic Analysis, Sentiment Analysis, Clustering, Most Used Sites For Crime.

1. Introduction
Now a day the use of social media is growing larger with the advancement of technology. This gives a scope of easy crime on the internet. This crime includes hacking of different accounts, make use of different fake profiles for blackmailing and demanding money from the people. So crime data mining came into existence. Crime data mining includes taking of data from different social media websites and comparing it with the crime data. If any there is any similarity between crime data and the text in the social media than there is a possibility of crime on that website. This helps in maintaining the law in that website as we know that there is a possibility of crime on that website. So, we can find some methods to avoid crime in the websites[1-3]. Sentiment analysis is the process of finding the most commonly used site or module in a site by the people in a browser this helps the browser to understand the mood of the user and act according to it. If a user uses a website for many times that he likes to use it in this way the sentiment analysis is done.
In this paper, we are combining the crime data mining concept and sentiment analysis to find the websites which are more affected by the crimes. To achieve this first we will find the websites in which there is a possibility of crime through crime data mining techniques. These websites which are the results of crime data mining will become the Dataset for sentiment analysis. The procedure of this analysis is discussed in this paper.

2. Crime data mining
The customary information mining strategies simply arrange the examples in organized information, for instance, order and fore-
3. Sentiment Analysis of Web Documents:

content notion investigation into bunch examination for organizing
the flow of online groups There is a collection of measure-
ments to set up web data, which include themes, systems, creators,
time et cetera. Content characterization in view of its passionate
extremity has changed into a currently developed desolate tract
that as it can, such sporting activities can sometimes
interrupt discovery to build up the affiliation policies from
clients' communiqué records. Specialists additionally can practice
this gadget to set up interlopers' profiles to assist discover capabil-
ity destiny system assaults. Like association manipulate mining,
consecutive instance mining reveals as frequently as feasible go-
ing on groupings of factors over an arrangement of exchanges that
befell underneath various occasions. In prepare interruption region,
this technique can apprehend interruption designs among time-
stamped facts. Indicating shrouded designs blessings wrongdoing
exam, but to gather widespread consequences calls for rich and
exceptionally prepared data[2].

3. Deviation recognition makes use of the precise measures to
think about information that varies recognizably from whatever is
left of the facts. Additionally, known as anomaly identification,
specialists can apply this strategy to extortion vicinity, arrange
interruption discovery, and other wrongdoing examinations. Be
that as it is able to, such sporting activities can some of the time
appear, by using all debts, to be regular, making it tough to under-
stand exceptions[3].

4. Characterization finds commonplace properties among one of a
kind wrongdoing substance and orchestrates them into predefined
instructions which have been linked for distinguishing the well-
spring of e-mail spamming as in keeping with the sender's auxilia-
ry highlights and phonetic examples. Frequently used to anticipate
wrongdoing patterns, grouping can decrease the time required to
distinguish wrongdoing materials. Nonetheless, the approach calls
for a predefined grouping plan. Arrangement likewise requires
sensibly entire getting ready and trying out statistics seeing that a
high degree of lacking facts would restriction expectation preci-
sion[4].

5. String comparator methods that demonstrate the connection the
literary fields in sets of database facts and parent the correspond-
ence the various statistics that can recognize tricky facts in crook
data, for example, the name and address. The scientists can use
string comparators to assess literary information that often calls
for escalated calculation. String correlation is the interesting area
for PC researchers that inside the case of string coordinating or
string separation measures. Levenshtein signify an ordinary mea-
ure of comparison among two strings as "alter do away with" along
those strains, the bottom quantity of, cancellations, unmar-
rried person inclinations, and substitutions need to alternate one
string into the alternative[5].

A portrayal of the hubs part of an inexpensive gadget is Social
machine investigation. Specialists can utilize this technique to broad-
en a system that represents culprits' elements, the circulate of
unmistakable and impalpable products and records, and the rela-
tionship amongst these factors. Encourage exam can uncover basic
components and subgroups and vulnerabilities within the network.
This technique empowers belief of crook systems; however, re-
tailers nevertheless won't have the capability to find the device's
real pioneers in the occasion that they live underneath the radar

C, you can utilize a web searcher on-line, for instance, Google,
and shoot the inquiry "C". It is beneficial to apprehend what a part
of the suits Google returns prescribes C as a motion goal. Consol-
idating assumption research into net searcher and content material
recovery advances empowers a more productive and sensible ad-
ministration for clients. Feeling examination has been utilized in
applications, as an example, information following and abridging,
on-line gatherings, document sharing, talking rooms, blogging and
so forth. YouTube provided assessment order innovation early this
12 months to kind every one of its feedback into "Poor" or "Great".
As a promising examination location, content material slant re-
search has been widely pondered, where evaluation examination is
utilized for content association assignments. [6-7] Existing evalua-
tion figuring techniques fall into two types: device getting to know
primarily based approach and semantic advantaged totally tech-
ique. Dialects which have been taken into consideration include
English. Our exploration means to additionally develop the soft-
ware.

4. Models and Methodology

Our method is for the most element crafted from the accompany-
ing advances: facts accumulation and purging, content end count
and stamping. Fig. 1 delineates the theoretical define of our tech-
nique, where three modules are characterized through incorporat-
ing content material end calculation. Module 1 is to alternate over
esteem-based totally information through content material feeling
calculation and investigation. In this module, any other catch-
phrase-based approach is familiar with ascertaining the belief
esteem for each little bit of content material by way of utilization of
the enterprise Java library created by Lieu Enterprise Search
and the How Net dictionaries. Our technique will yield various
an incentive for every publish, with the sign demonstrating its enthu-
siastic extremity and the outright esteem its passionate pressure.

![Block Diagram of actual work](image-url)

**Fig. 1: Block Diagram of actual work**

4.1. Data collection:

Before data slithering and purging manner are commenced, an in-
depth perspective of the shape of Sinaa sports activities organiza-
tion is vital. Online Sinaa sports group shows a tree-like shape
with root discussions, department gatherings and a no separable
base layer of leaf gatherings. There are altogether forty-nine leaf
gatherings for this group. Fig. 2 outlines the tree-like structure of

---

International Journal of Engineering & Technology

---

63
the Sinaa sports institution, where the foundation hub, pink circle hub, and yellow rectangular hub communicate to the complete institution, the number one-layer gatherings, and the leaf discussions one at a time.

We continue with the records slithering and purifying technique inside the accompanying four levels:

Stage 1: Physically make a table
SINAA1_LEAFORUM1_URLLIST
In progression, we physically store the data for all the 49 gatherings into a table named SINAA1_LEAFORUM1_URLLIST in the database, where their names and URL joins are available.

Stage 2. Make table SINAA1_FORUM1_URL1 in view of SINAA1_LEAFORUM1_URLLIST after the obtaining of the connections for all the leaf gatherings, we parse the principal pages of them inside and out and produce a rundown of URLs of site pages that contain all the point posts and the remark posts. The rundown will be built into the SINAA1_FORUM1_URL11 table in the database.

Stage 3. Cross the connections in the SINAA1_FORUM1_URL1 table and slither down every one of the presents This progression is on navigating through every one of the connections that are in the SINAA1_FORUM1_URL1 table, to parse out all the point and remark posts contained on the relating pages, and to store them into two tables of SINAA1_FORUM1_TOPIC1_POST1 and SINAA1_FORUM1_COMMENTS1_POST1. Two parsing layouts are composed in XML arrangement to parse the posts, which are Sina1SportForumReplyPostParseTemplate.xml and Sina1SportForum TopicPostParseTemplate.xml.

Fig. 3 shows the creeping methodology and the structure of the social tables and the XML layouts, where the green featured thing in the tables are the essential keys.

5. Computation of content sentiment from Forum Posts:

In this segment, semantic creation-based approach may be produced utilizing any other calculation by consisting of the perception esteem for every single watchword to perform the notion esteem for the entire article. Content end examination is gone for computing an entire wide variety an incentive for every bit of text, the full estimation of which speaks to the persuasive strength and the indication of which indicates its enthusiastic extremity. Assume the present publish is p seeing that it is composed in Chinese, we to begin with use PC based programmed phrase division equipment to interrupt p into an expansion of watchwords w1, w2, w3, ..., wn, in which there are n of them altogether. Each catchphrase wi(i=1,2,3,..., n) could be allowed an opinion esteem vi by our proposed calculation
while the assumption esteem for p is the total of the opinion esteeems for all the catchphrases. Give \( V_p \) a chance to indicate the conclusion esteem for p and we have
\[
V_p = \sum_{i=1}^{n} v_i
\]
Figuring of the opinion esteem cluster \{v1, v2, v3, ..., vn\} depends on catchphrases examination and coordinating. With a specific end goal to compute the assessment esteem for each watchword contained in p.

6. Conclusion

Here in this paper, we developed a method for detecting most crime sited websites to this we used clustering algorithm in crime data mining and mathematical expression for sentiment analysis. Through this method, we can control or secure that website through providing security once it is implemented.

References

[1] K. Ahmad, Y. Almas, Visualising sentiments in financial texts? Proceedings of the Ninth International Conference on Information Visualisation (2005) 363–368.
[2] C. Asavathiratham, The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains, Dept. of EECS. 2000, MIT, Cambridge, 2000, p. 188.
[3] P. Chaovalit, L. Zhou, Movie review mining: a comparison between supervised and unsupervised classification approaches, Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.
[4] K.W. Cheung, I.T. Kwok, M.H. Law, K.C. Tsui, Mining customer product ratings for personalized marketing, Decision Support Systems 35 (2) (2003) 231–243.
[5] J. Coble, D. Cook, R. Rath, L. Holder, Iterative structure discovery in graph-based data, International Journal of Artificial Intelligence Techniques 1–2 (14) (2005) 101–124.
[6] Fayyad, U.M., and Uthurusamy, R. ( Aug. 2002). Evolving Data Mining into Solutions for Insights. Comm. ACM. 28-31.
[7] Hosseinkhani, J., Chuprat, S., and Taherdoost, H., (2012a). Criminal Network Mining by WebStructure and Content Mining. 11th WSEAS International Conference on Information Security and Privacy (ISP ’12), Prague, Czech Republic September 24-26.
[8] Hosseinkhani, J., Chuprat, S., Taherdoost, H., and Shahraki Moghaddam, Amin., (2012b). Propose a Framework for Criminal Mining by Web Structure and Content Mining. International Journal of Advanced Computer Science and Information Technology (IJACSIT), Helvetic Editions. 1(1). 1-13.
[9] Keyvanpour, M., Javiedeh, M., and Ebrahimi, M. (2011). Detecting and investigating crime by means of data mining: a general crime matching framework, Elsevier, Procedia Computer Science 3 (2011) 872–880.