**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 6, No. 3, June 2016, pp. 995 ~ 1001
ISSN: 2088-8708, DOI: 10.11591/ijece.v6i3.9878 

# Automatic Detection of Illegitimate Websites with Mutual Clustering

**K. Kanaka Durga, V. Rama Krishna**
Dept of CSE, K L University, Guntur, AP

**Article Info**

***Article history:***

Received Jan 6, 2016
Revised Mar 10, 2016
Accepted Mar 25, 2016

***Keyword:***

Illegitimate
Mutual Clustering
Phising
Web Crawled

**ABSTRACT**

In the websites the contents will be are similarity when we compared with other search engines. So to check the similar content in the websites and its web contents we created a overhead to the search engine which will severely effect its performance & quality. So to detect the silmilar or same content or web documenattion some techniques are implemented by web crawling research community. So it is one of major factor for the search engines to provide some applicatory data to users in the first page itself. So to avoid such issues we proposed a methodlogy called Automatic Detection of illegitimate websites with Mutual Clustering (ADIWMC) paper we are presenting a peculiar and efficacious path for the detection of similarities in the web pages in web clustering. Detection of same and similar web pages and web content will be done by storing the crawled web pages into depository. Initially the adwords will be extracted from the crawled pages and similarity checking will be done between the two pages based in the usage of adwords. So a threshold value is set for this, if the similarity checking percentage is greater than the threshold then similarity content is reduced and improves the depositary and improves the search engine quality. In the sections of existing analysis and the proposed analysis we are clearly exploring how it works.



***Corresponding Author:***

K. Kanaka Durga,
M. Tech Student, Dept of CSE,
K L University,
Vaddeswaram 522502, Guntur District, Andhra Pradesh, India.

## 1. INTRODUCTION

Large-scale and targeted attacks: Cybercriminals have to cheat online strategy adopted two familiar consumer. Many scams are designed for large-scale success [1]. Phasing scams posing as banks and online service providers by the thousand ray's million spam messages fail fraction of users to a fake website penal control [2]. In fact, many thieves are working somewhere in between, faithfully reproduce the logic of fraud, without hardware to reproduce from previous versions of the attack. Thus, criminals engaged in advanced banking fraud cost places exist for banks with online banking, which the victim has access to the inspection of their 'deposits'. When a false bank is off, the criminals a new optimized from the old site. Criminals have the fake escrow services as part of an advanced higher tax fraud. On the surface, escrow sites seem different, but often share similarities in the text or HTML structure leg. Yet another example is online Ponzi 'high yield investment programs (HYIP). The programs offer investors the extravagant interest, which means that inevitably collapse when dry attract new deposits. The authors are behind the scenes as the creation of new programs that often share similarities with previous versions [3]. The designers of these scams have a strong incentive to distinguish from the old to keep their new copies. Potential victims may be afraid when they realize that an earlier version of this site, "reported as fraudulent. So, criminals a concerted effort to distinguish new copies of old.

The main goal of the research is to develop new and effective detection of the same name near the site documents. In the beginning of the crawl web pages must be prepared with the help of which the data parsing HTML tags and Java web writing, remove the documents [4]-[8]. This is followed with the removal of the legal form or the end of the crawled pages. And the effects of the Algorithm are used to filter the affixes (prefixes and suffixes) to crawl data for the keywords. Finally, the similarities between the two documents number is calculated based on the extracted keywords. The documents are received at the difference is greater than a predetermined threshold value in the neighborhood of the same name. Much research has been done many experiments using real data sets, and has been found to surpass the previous recommendation algorithms.

## 2. LITERATURE SURVEY

Literature Research is the most important step in the development of software. Before the new tool should specify the reasons for the company's economic power. Once these things are complete, the following steps to determine which operating system and could be used for the development of the device [9]. If the work begun construction software applications need much outside support. Supporting this software can be obtained from above, from the book or the website. For the construction of the system of the above consideration the development of the system was taken into account.

### 2.1. Existing Methodology

The two main types of vines are common and on the skin. Previous documents and links to various factors common to the crawler to crawl, but it was in the past as a way to limit the number of pages with the help of some special knowledge has been focused crawler. Repositories index of the Web page, to me, and not the search page on a system (for example, Research) was established by the web crawler for the offer advice [10]. The development of the Internet in order to survive a close copy of the document and with the need to integrate heterogeneous data serious problems. They also bear a striking similarity among the dummy data is not the same a little bit closer. Web research is facing serious problems because of the duplicate copies and is close to the web page. Page or index storage location or irritate users to slowly increase or increase the cost of this service. Therefore algorithms that recognize the inevitability of such a page [11]. Web crawling issues such as freshness and effective use of the resources referred to in the past. Recently, the elimination of duplicate or near duplicate documents on the Internet has become a major concern and has attracted significant research. The survey is beyond the boundaries of the significance of the door handle cyber crimes, as each will be very small, enforcing the law could prove to be a valuable way. For example, the method in two different groups, each of which explains more than 100 fake escrow sites. In addition, you can reduce the workload for investigators, according to the priorities that will investigate how criminals. We have many promises to continue the kind of work you. [12] First, the true Yip clustering is good. Secondly, it is a sign of clustering phishing sites and spam ads in storefronts, as has been tried in other areas where it would be interesting to compare the combination of clustering. Finally, additional input features such as Whois registration and image detail.

## 3. PROPOSED SYSTEM

Keywords fake documents collected from the breast close to the web. First, crawled web documents are parsed to extract the keywords. Parse / common language to stop and view the rest of the term, HTML tags are JavaScript removal. And to reduce the number of keywords that closely matches the name plate on the table is stored in the process. The password is stored in the table so that the search space is reduced to the breast. Against all the files in the repository of an equal number of documents on the Internet today are computed from a list of keywords on the page. Documents before the similarity score is considered to be close to the threshold of the same name more than.

In this paper, we have a history of close and effective Web crawling the site to identify the duplicate post. Catch the close of the Web page in advance to reserve a copy of the web pages have been crawled repositories [13]. At first from crawling pages, keywords, keywords are extracted and collected score is calculated based on the similarity between the two pages. The document with the same name is considered to be particularly important as the closing of the door is more than many similarities.

### 3.1. Study of system & Appropriate Implementation

The appropriate implementation of the project and business process to examine this proposal and cost estimates for the project and will have to put forth some of the best overall plan. The implementation of a system analysis of the proposed system at the time of the trial was conducted appropriately. This is to

ensure that the proposed system will not be a burden for the company. Some of the more understanding of the circumstances necessary for the implementation of the appropriate method of analysis. The three most important are interested in participating in the analysis of the appropriate performance are as follows:

**a. Economic appropriate enforcement**

The company's structure is designed to ensure that the economic impact of this study. The system can pour into the company's research and development is limited to the amount of the subsidy. The costs have to be justified. The majority of the budget within the framework of the development and use of technology, it is possible for free. Should purchase only products manufactured.

**b. Technical appropriate enforcement**

In this study, the appropriate technical implementation, the system is designed to verify the means of the technical requirements. The system to be developed without a high demand for technical resources. These technical resources will lead to higher demands. This will lead to high demand placed on the client. Development of the system necessary for the application of such property or vacuum the system measures only modest changes, must have.

**c. Social sufficient to run**

The sample of the study to verify the level of user acceptance of the system. Including user training to use the system effectively in the process. The system user should not feel threatened, but we need to accept it. At the height of the agreement is the only user of the system depends on the method of consumer education and to do good works. The level of confidence are the end users of the system, welcomed some useful criticisms will not be able to be built.

## 4. SYSTEM ARCHITECTURE

### 4.1. Installation & Implementation

Project implementation process transforms theory in the operating system. Therefore, it is considered the most important sector in the effective implementation of the new system, giving the user confidence that the new system will work and be effective. The establishment of good planning process including analysis of existing systems and the application of pressure, made way for the passage and the evaluation of the transition (Figure 1).

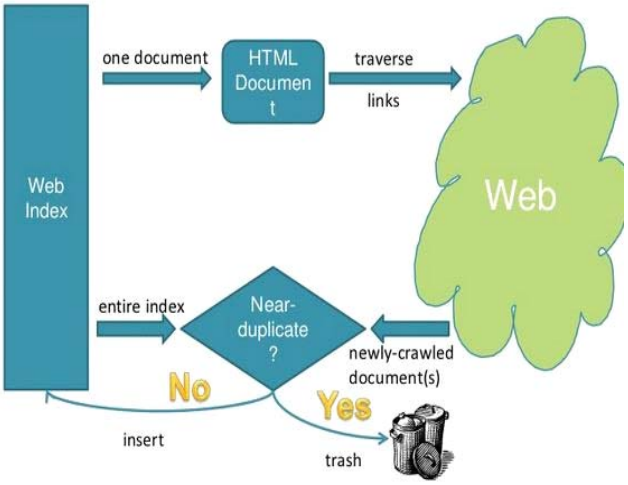


Figure 1. System Architecture

**a. Web document parsing**

Information extracted data to crawl to help determine the future path of the track. The analysis can be as simple as a / hypertext link or URL extraction cleaning difficult as the analysis of HTML tags in the HTML content. Is inevitable for the analyzer, which is available on the whole site will be faced with many

errors. The analyzer usually get information from the Web page can not think of a few words, as usual, and many other HTML tags, JavaScript and other bad characters (Figure 2).

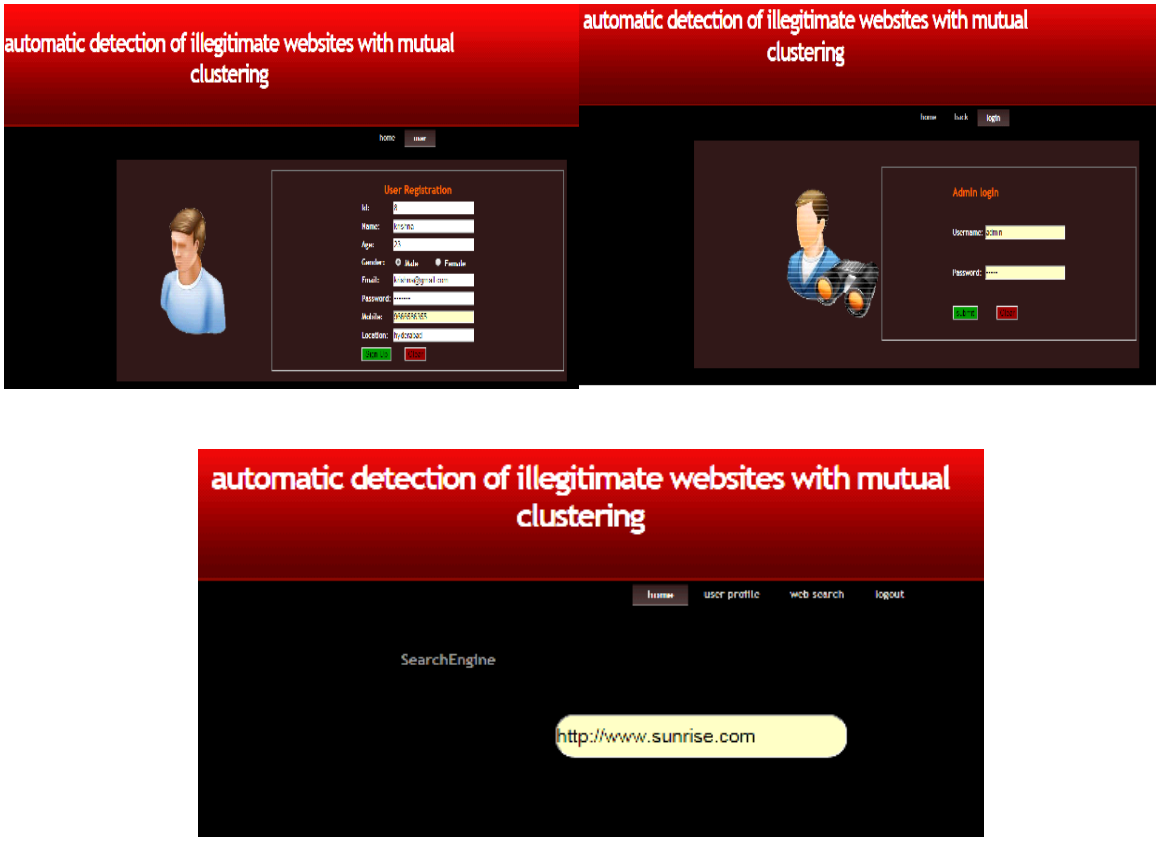


Figure 2. Entering the website in search engine

**b. In view of the algorithm**

The information recovery variants are limited to a common root by reference. After the false premise that the two share the same possess profound expression. Thus, they have the same ideas still appear morphologically different is known in the IR system using the same query and the document should be checked using the view [14]. Given facilitate the reduction of words that have the same root as. This is accomplished through the removal of each derivative and inflectional s+uffixes. [15] For example, is "connected", "connected" and "communication" to all condensed to "connect".

**c. Keywords representatives**

We have a unique password and is counted each time they crawl a web page as a result of view. These keywords represent the image to improve the process closely as possible. This is the picture in order to reduce the search space for detecting near duplicate. First of keywords and the number appear on the web pages are arranged in the order based on the descending census. After that, the number N of the keyword number is stored in the table, and the remainder is indexed keywords and stored in another table. In this way, the value of N is set to 4. The difference between the two numbers on documents can only be calculated if the first two keywords of the document are the same (Figure 3). Therefore, the search area is reduced to close in the same name.


Figure 3. Checking with other keywords

### d. Contrast Score Calculation

If the first key word in the new web page is not in compliance with the first keywords for the page in the table, the information added to the web page in the repository. If all the keywords of the two pages is the same information should be considered duplicate pages and therefore are not included in the repository. If the first new keywords on the same page in the repository, and the similarities between the two documents can calculate numbers (Table 1).

Table 1. Internet equal to the sum of these two documents is calculated

| | | | | | | |
|---|---|---|---|---|---|---|
| T1 | K1 | K2 | K4 | K5 | …… | Kn |
| | C1 | C2 | C4 | C5 | …… | Cn |
| T2 | K1 | K3 | K2 | K4 | …… | Kn |
| | C1 | C3 | C2 | C4 | …… | Cn |

Let T1 and T2 for tablets containing the extracted keywords and the corresponding importance. The label of each table is regarded as the calculation of the similarity score. If a label is present in both tables, the formula used to calculate the tag similarity score is as Figure 4 and Figure 5.

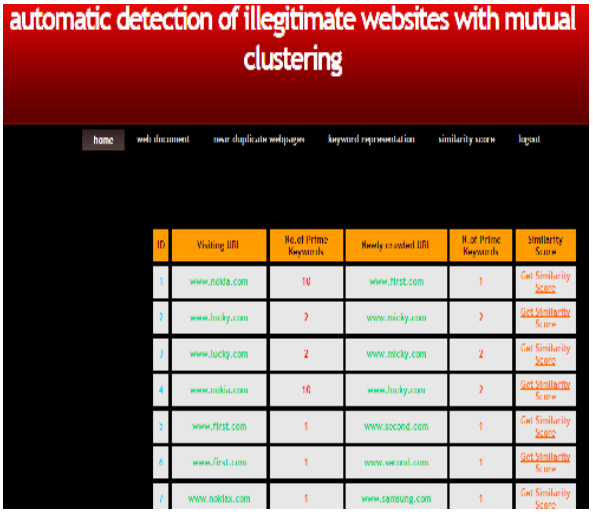


Figure 4. (A) Exact Result analysis of Content similarities in websites

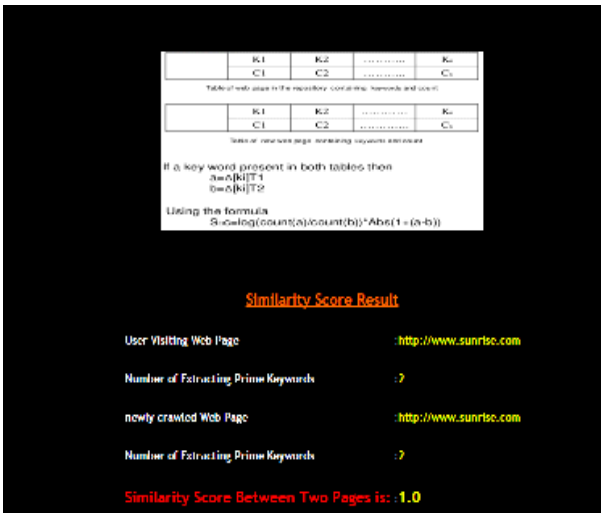


Figure 5. (B) Examining in Percentage of Similarity

**e. Web document duplicate detection close**

Near the same name will not be considered the "real names" which documents the minute differences. Typographical errors, release, radiation, or plagiarized documents, many expressions of the same physical things, spam emails from the same naked form, and a number of these cases may result in a heap Reaching almost similar. The percent of the large web page is known to be close to a variety of the same name, according to the survey. These surveys suggest that as part of close to 1.7% to 7% of the web page to go crawler. The steps involved in the method outlined in the following paragraphs (Figure 6).

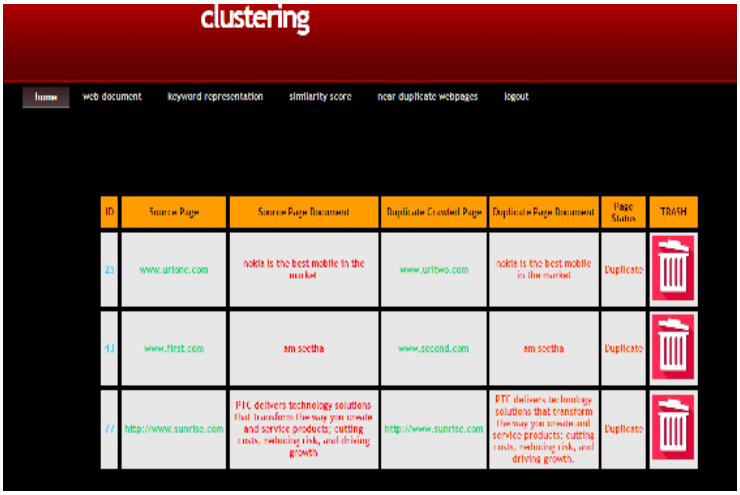


Figure 6. Deleting the Duplicate Content

An output is a quality that helps the end user needs and presents information clearly. Be in any system of treatment results communicated through the vents to users and other systems. The design capacity is determined how the information needed immediately and must be moved to the output copy paper. It is the main source of information and direct user. The efficient and intelligent output design increases the ratio of the system to help the user in decision making.

## 5. CONS OF THE SYSTEM

- Design computer production must take place in an organized, well thought so; law, the output to be developed with the proviso that each output is designed so that the system is found to be convenient and efficient to use. When the output of the analysis of the computer designs, they are to the specific output which is needed to identify the requirements.
- Select methods for presenting information.
- Create document, report or other formats that information contained in the system.

The output form of an information system needs to perform one or more of the following objectives. Send information about past, current and projected state of the future. Report of significant events, opportunities, problems, or warnings and trigger action to confirm an action.

## 6. CONCLUSION

In this paper we explored that the web consists of information/data storing in various features inlcudes such as using of structured and unstructured data which exbits their dynamic nature and checking the similaritiy in the website contents which will effect the data retriveal. Then the web page documents have pose high level difficulties for the search engines to know which websites are of real or fake. So the detection of similarity and same content check will gained more attention in present years using web mining techniques. So in this paper we propose a new methodology which combine and evaluate the clustered methodolgy to automatic link which gets the semi-automatic scams. By the simulation and evalution results we have shown the more accuracy rate than the GPC clustered approach. It can also use in big scams like phishing that use more copy of content. Particularly we applied this for two scams called HYIPS and fake escrow websites. Further work can be extented to the reduced memory spaces for more web depositaries and which improves the search engine quality.

## REFERENCES

[1] S. Acharya, *et al.*, "Selectivity estimation in spatial databases," in *SIGMOD*, pp. 13–24, 1999.
[2] S. Alsubaiee, *et al.*, "Supporting location-based approximate-keyword queries," in *GIS*, pp. 61–70, 2010.
[3] A. Arasu, *et al.*, "Incorporating string transformations in record matching," in *SIGMOD*, pp. 1231– 1234, 2008.
[4] A. Arasu, *et al.*, "Efficient exact set-similarity joins," in *VLDB*, pp. 918–929, 2006.
[5] N. Beckmann, *et al.*, "The R tree: an efficient and robust access method for points and rectangles," in *SIGMOD*, pp. 322–331, 1990.
[6] A. Z. Broder, *et al.*, "Min- wise independent permutations (extended abstract)," in *STOC*, pp. 327–336, 1998.
[7] X. Cao, *et al.*, "Retrieving top-k prestige-based relevant spatial web objects," *Proc. VLDB Endow.*, vol. 3, pp. 373–384, 2010.
[8] K. Chakrabarti, *et al.*, "An efficient filter for approximate membership checking," in *SIGMOD*, pp. 805–818, 2008.
[9] S. Chaudhuri, *et al.*, "Robust and efficient fuzzy match for online data cleaning," in *SIGMOD*, pp. 313– 324, 2003.
[10] S. Chaudhuri, *et al.*, "Selectivity estimation for string predicates: Overcoming the underestimation problem," in *ICDE*, pp. 227–238, 2004.
[11] S. Chaudhuri, *et al.*, "A primitive operator for similarity joins in data cleaning," in *ICDE*, pp. 5–16, 2006.
[12] E. Cohen, "Size-estimation framework with applications to transitive closure and reachability," *Journal of Computer and System Sciences*, vol/issue: 55(3), pp. 441–453, 1997.
[13] G. Cong, *et al.*, "Efficient retrieval of the top-k most relevant spatial web objects," *PVLDB*, vol/issue: 2(1), pp. 337–348, 2009.
[14] G. Li, *et al.*, "Supporting search-as-you-type using sql in ~databases," *TKDE*, 2011.
[15] A. Mazeika, *et al.*, "Estimating the selectivity of approximate string queries," *ACM TODS*, vol/issue: 32(2), pp. 12–52, 2007.