ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Reusability Testing for Code Clone Detection in Web Applications

Edith Linda P[1], Sasikala K[2]

Assistant Professor, School of IT and Science, Dr.G.R.Damodaran College of Science, Tamilnadu, India[1]

Research Scholar, School of IT and Science, Dr.G.R.Damodran College of Science, Tamilnadu, India[2]

ABSTRACT: Code clone is a process of detection and analyzing the web pages enabled to accumulate data to boost the quality and conceptual/design of the info of the online application. The Proposed work uses Levinstein Edit Distance Algorithm to find the code clones present in the static and dynamic web applications by Computing the similarity percentage, which is based on the observation that the reserve a matrix to hold the Levinstein distances between all prefixes of the parsed code and web tags and all prefixes of the second parsed code and web tags, then to compute the values in the matrix in a dynamic programming , and thus find the distance between the two web pages matrix values are computed.Source code reusability testing that associates an approach to clone detection and analysis for dynamic web applications has been planned at the side of a paradigm implementation for sites. The similarity degree is often custom-made and tuned in an exceedingly easy approach for various net applications. The results have confirmed that the termination of study and style of the online application has result on the duplication of the pages; these results allowed testing companies to spot some common options and therefore the collocated workshops that would be integrated, by deleting the duplications.

KEYWORDS: Code Clone, Reusability, Duplicate pages.

I. INTRODUCTION

Source code reusability is an approach that automatically tests and refactors (testing and refactoring) and it detects the duplicated pages in modern web applications in dynamic websites. And it analysis for both the page structures, developed by specific sequences of HTML tags, and the displayed content Actually, the system attempts to exploit the results of the code clones present in static and dynamic web pages by applying Levinstein Edit Distance Algorithm to support dynamic web application reengineering activities. The approach is to analyze the page structure, enforced by specific sequences of markup language tags, and therefore the content displayed for each dynamic and static pages.[9] Moreover, for a combine of sites the system tends to conjointly think about the similarity degree of their similarity percentage. The detection and analysis of the web page code enabled to acquire information to improve the general quality and conceptual/design of the code.

II. OVERVIEW OF THE AREA

Code clones square measure similar program structures of significant size and vital similarity. Many studies recommend that the maximum amount as 20-50% of many software system systems include cloned code. Knowing the placement of clones helps in program understanding and maintenance [5]. Some clones is removed with refactoring, with a large number of clone detection techniques been planned within the literature. One limitation of the present analysis on code clones is that it's largely targeted on the fragments of duplicated code and not observing the large image wherever these fragments of duplicated code square measure probably a part of a much bigger replicated program structure. The

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

approach is based on the computation of the LE distance metric. The built-in ASP objects, together with their methods, properties and collections, may characterize the control component of an ASP page. Text clustering which directly related to the concept of data clustering. Document clustering is a more definite technique for unsupervised document organization, automatic topic extraction and fast system analysis or filtering. Other approaches address the problem of combining together finite documents in business environments, for instance separating business letters from technical papers in last few years the classification of pages in journals and books received more attention. An similar aspect of page classification are the features that are extracted from the page and used as input to the classifier. Sub-symbolic features, refers density of black pixels in a region, region computed directly from the image.

III. OBJECTIVES OF THE ANALYSIS

The system proposes associate approach to automatically test and refactor modern net applications and detect duplicated pages in dynamic web sites. An analysis of each page structure, enforced by specific sequences of hypertext mark-up language tags, and they are displayed as graphical representation. Additionally, for every combination of dynamic pages the system has a tendency to define the similarity degree of their scripting code. The similarity degree of two pages is completely computed and the different similarity metrics for the various components of a web page supported the code duplication string and they edited. The system have enforced a model to automatize the clone detection method on net applications to develop the technology and used to validate our approach in an exceedingly case study.

IV. LITERATURE SURVEY

In this chapter, describes a comprehensive survey and literature review of existing work. All the review of literature is related to webpage extraction and they are focused only in desktop application. The chapter 2 contains different types of webpage code extraction journal details, all these journals are related to code clone analysis. Here, how the Web pages are being extracted and how they are used for the purpose of code clones and the algorithms used in their works are explained.

Ohta, Takafumi, et al, [7] proposes a "Source code reuse evaluation by using real/potential copy and paste." It is used for a webpage to detect the clones and calculates the majority of codes that are similar to each other. This increases the software quality. It is only used in desktop applications. Algorithm used here is Levenshtein distance algorithm.
ArvindArasu, Hector Garcia-Molina**,** [1] proposes an Extracting organized information from Web pages Many sites contain extensive arrangements of pages created utilizing a typical format or design. For instance, Amazon lays out the creator, title, remarks, and so on in the same path in all its book pages. The qualities used to produce the pages regularly originate from a database.In this paper, they concentrate on the issue of naturally separating the database values from such format produced website pages with no learning samples or other comparative human information. The framework introduce a calculation that takes, as data, an arrangement of format created pages, reasons the obscure layout used to produce the pages, and concentrates, as yield, the qualities encoded in the pages. Test assessment on countless information page accumulations shows that our calculation effectively separates information much of the time Algorithm used Markov Decision Processes.
Chang, Chia-Hui, and Shao-Chen Lui, [2] proposes a “IEPAD: Information Extraction Based on Pattern Discovery,” The examination in data extraction respects the era of wrappers that can remove specific data from semi organized Web records. Like compiler era, the extractor is really a driver system, which is gone with the created extraction standard. Past work in this field means to take in extraction rules from clients' preparation sample. Here the framework proposes IEPAD, a framework that consequently finds extraction rules from Web pages. This new track to IE includes no human exertion and substance subordinate heuristics. Test results demonstrate that the developed extraction guidelines can accomplish 97 percent extraction more than fourteen mainstream internet searchers Algorithm used is PAT(Patricia tree) trees Algorithm

Chang, Chia-Hui, et al, [3] Proposes the “Survey of Web Information Extraction Systems,” Web displays a tremendous measure of valuable data which is generally organized for its clients, which makes it hard to separate pertinent information from different sources. Subsequently, the accessibility of strong, adaptable data extraction (IE) frameworks that change the Web pages into system inviting structures, for example, a social database will turn into an extraordinary

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

need. Numerous methodologies for information extraction from Web pages have been developed, there has been restricted push to look at such apparatuses. The framework overviews the significant Web information extraction methodologies and thinks about them in three measurements: the assignment area, the computerization degree, and the strategies utilized. The criteria of the first measurement clarify why an IE framework neglects to handle some Web destinations of specific structures Algorithm used here is Extended Co-Citation Algorithm.

Hsu, Chun-Nan, and Ming-Tzung Dung, [4] propose an"Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web," Incorporating countless data sources might altogether expand the utility of the World-Wide Web. A promising answer for the joining is through the utilization of a Web Information middle person that gives consistent, straightforward access for the customers. Data go betweens need wrappers to get to a Web source as an organized database, however assembling wrappers by hand is unrealistic. Past work on wrapper actuation is excessively prohibitive, making it impossible to handle a substantial number of Web pages that contain tuples with missing qualities, various qualities, variation property stages, exemptions and mistakes. The framework have executed this methodology into a model framework and tried it on genuine Web pages. The execution measurements demonstrates that the sizes of the affected wrappers and in addition the required preparing e ort are straight with respect to the basic change of the test pages Algorithm used here is FST(Finite State Transducer) Algorithm.

Laender, Alberto HF, et al, [6] proposes "A Brief Survey of Web Data Extraction Tools," In the most recent couple of years, a few works in the writing have tended to the issue of information extraction from Web pages. The significance of this issue gets from the way that, once extricated, the information can be taken care of in a path like occasions of a customary database. The methodologies proposed in the writing to address the issue of Web information extraction use procedures acquired from zones, for example, normal dialect preparing, dialects and syntaxes, machine learning, data recovery, databases, and ontology's. As an outcome, they show extremely particular components and abilities which make an immediate examination hard to be finished. In this paper, they propose a scientific categorization for portraying Web information extraction fools, quickly study real Web information extraction devices depicted in the writing, and give a subjective examination of them. Ideally, this work will invigorate different studies went for a more complete examination of information extraction methodologies and apparatuses for Web information. Algorithm used here is hyperlink analysis-based algorithm.

V. CODE REUSABILITY

The other name for Code reuse is software reuse. It is used in existing software or software knowledge to build new software. Code reuse aims to save time and resources and reduce redundancy by taking advantage of assets that have already been created in some form within the software product development process. The key idea in reuse is that parts of a computer program written at one time can be or should be used in the construction of other programs written at a later time. Useful ways to develop cost effective software, especially with the availability of huge amounts of open-source work [8]. Reuse saves cost, increases the speed of development and improves software reliability. However, the quality of less known packages and the large number of projects developed by programming enthusiasts is unknown. Reusing them may be the source of more problems rather than being a solution to a problem. The reuse of source code within sections of an application and potentially across multiple applications. At its best code reuse is accomplished through the sharing of common classes and/or collections of functions and procedures. At its lowest code reuse is accomplished by copy, paste, and the result is negative overall value.

VI. PROPOSED WORK FOR DETECTING CLONES

The proposed work initially extracts a HTML page to the local and parsing the item gives one token at once, much as a document handle gives you one line at once from a record. The HTML can be tokenized from a record or string. A system that concentrates data by working with a flood of tokens doesn't need to stress over the eccentricity of substance encoding, whitespace, quotes, and attempting to work out where a label closes. Customary expressions are effective, yet they are extremely low-level method for managing HTML. The framework forms the spaces and new lines, single and a copy cites HTML remarks, and significantly more. The following step up from a customary expression is a HTML tokenize. In this part, the HTML Parser is utilized to extract the data from HTML documents. Utilizing these

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

systems, you can extricate data from any HTML record, and never again need to stress over character-level trivia of HTML markup. The framework performs an intriguing trial on finding dependable sites. It is understood that Google (or other web indexes) is great at finding legitimate sites. Be that as it may, do these sites give precise data. To answer this inquiry, the framework looks at the online site pages that are given most noteworthy positions by Google with the pages with most noteworthy reliability found by Levenshtein separation utilizes iterative techniques to figure the site dependability and certainty, which is generally utilized as a part of numerous connection examination approaches. The basic element of these methodologies is that they begin from some starting express that is either arbitrary or uninformative. The confinement of considering just straightforward clones is known in the field. The fundamental issue is the tremendous number of basic clones commonly reported by clone recognition instruments. There have been various endeavors to move past the crude information of straightforward clones. It has been proposed to apply grouping, sifting, representation, and route to offer the client some assistance with making feeling of the cloning data. Another path is to identify clones of bigger granularity than code pieces.

VII. LEVENSHTEIN DISTANCE ALGORITHM FOR CODE CLONE DETECTION

Levenshtein distance algorithm might be a string metric for measuring the amount of difference between two sequences. In approximate string matching, the idea is to find matches for short strings, for particular strings from a vocabulary, in many longer texts, in situations where a small number of variations is to be predictable. Here, one of the strings is normally short, whereas the other is arbitrarily long. This has a wide range of applications, for instance, spell identifier, correction systems for optical character recognition, and software to support natural language translation based on translation memory. The Levenshtein distance is able to compute between two longer strings, but the rate to compute it that is roughly proportional to the product of the two string lengths makes this impractical. Thus, once used to aid in fuzzy string searching in applications such as record linkage, the compared strings square measure usually short to help improve speed of comparisons.

VIII. RESULTS AND DISCUSSIONS

CODE clones are related program structures of significant size and significant similarity. Several studies suggest that the maximum amount as 20-50 percent of large software systems contains cloned code. Knowing the location of clones helps in program understanding and maintenance. Some clones are often removed with refactoring, by replacing them with function calls or macros, or the system can use unconventional meta level techniques such as Aspect-Oriented Programming to avoid the harmful effects of clones. Cloning is an active area of research, with a large number of clone detection techniques been proposed in the literature. One limitation of the current research on code clones is that it is mostly focused on the fragments of duplicated code (we call them simple clones), and not looking at the massive image where these fragments of duplicated code are possibly a part of a bigger replicated program structure. These larger roughness similarities are called as structural clones**.**

IX. CONCLUSION AND FUTURE WORK

Refactoring modern web applications and defines duplicated pages in dynamic web pages an automated reusability testing will be implement with the help of source code reusability testing.

- A series of fault models that may be mechanically checked on any program state, capturing completely different categories of errors that area unit doubtless to occur in ASP net applications
- Given the growing quality of ASP net applications, we have a tendency to see several opportunities for exploitation in apply
- The open supply and plug-in-based nature makes our tool an appropriate vehicle for different researchers curious about experimenting with different new techniques for testing ASP net applications

Future analysis on net knowledge extraction focuses on comparison the contents showing on the page still because the code to live the quality and originality of the online page. However, they are redesigned or applied in an exceedingly completely different sequence and state of affairs to resolve key problems in page-level knowledge extraction and comparison to the code of computer and its contents to seek out the pretend and also the real. The System also can

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

worked to observe the script injection and projected towards the detection of malwares connected to web content, that harms the users machine and acts as a spy ware and sends the knowledge of the top user to the aggressor. These systems are still in analysis to stop the attackers.

REFERENCES

1. Arasu, Arvind, and Hector Garcia-Molina. "Extracting structured data from web pages." Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, 2003.
2. Chang, Chia-Hui, and Shao-Chen Lui. "IEPAD: information extraction based on pattern discovery." Proceedings of the 10th international conference on World Wide Web. ACM, 2001
3. Chang, Chia-Hui, et al. "A survey of web information extraction systems."Knowledge and Data Engineering, IEEE Transactions on 18.10 (2006): 1411-1428.
4. Hsu, Chun-Nan, and Ming-Tzung Dung. "Generating finite-state transducers for semi-structured data extraction from the web." Information systems 23.8 (1998): 521-538
5. Khan, Muhammad Zahid, and M. N. A. Khan. "Enhancing Software Reusability through Value Based Software Repository." International Journal of Software Engineering and Its Applications 8.11 (2014): 75-88.
6. Laender, Alberto HF, et al. "A brief survey of web data extraction tools." ACM Sigmod Record 31.2 (2002): 84-93.
7. Ohta, Takafumi, et al. "Source code reuse evaluation by using real/potential copy and paste." Software Clones (IWSC), 2015 IEEE 9th International Workshop on. IEEE, 2015.
8. Rattan, Dhavleesh, Rajesh Bhatia, and Maninder Singh. "Software clone detection: A systematic review." Information and Software Technology 55.7 (2013): 1165-1199.
9. Tan, Chin Loong. "IMPLEMENTING A Web-based Computerized restaurant system."