A Ruse Secluded character set for the Source

Mr. J Purna Prakash[1], Assistant Professor Mr. M. Rama Raju [2], Assistant Professor
Christu Jyothi Institute of Technology & Science

Abstract

We are rich in data, but information is poor, typically world wide web and data streams. The effective and efficient analysis of data in which is different forms becomes a challenging task. Searching for knowledge to match the exact keyword is big task in Internet such as search engine. Now a days using Unicode Transform Format (UTF) is extended to UTF-16 and UTF-32. With helps to create more special characters how we want. China has GB 18030-character set. Less number of website are using ASCII format in china, recently. While searching some keyword we are unable get the exact webpage in search engine in top place. Issues in certain we face this problem in results announcement, notifications, latest news, latest products released. Mainly on government websites are not shown in the front page.

To avoid this trap from common people, we require special character set to match the exact unique keyword. Most of the keywords are encoded with the ASCII format. While searching keyword called cbse net *results thousands of websites will have the common keyword as* cbse net *results. Matching the keyword, it is already encoded in all website as ASCII format. Most of the government websites will not offer search engine optimization. Match a unique keyword in government, banking, Institutes, Online exam purpose. Proposals is to create a character set from A to Z and a to z, for the purpose of data cleaning. Example Indian rupee has unique code is 20b9 (₹) for the symbol is encoded. Make private use code point. Same way using UTF-8, UTF-16, UTF-32. can approach banking, official websites, e-commerce, online security portals*. Getting all platforms into one roof.

Keyword: UTF, Font, keyword Matching, Encoding, character set, Allograph.

I. INTRODUCTION

Unicode is an entirely new idea, to create a binary code for text, it is also called Unicode worldwide character standard. It works as Interchange, processing and display of the written texts.at present Unicode standard contains 34,168 distinct coded characters are derived from language scripts.

Fundamentally, computers just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different systems, called character encodings, for assigning these numbers. These early character encodings were limited and could not contain enough characters to cover all the world's symbols. Even for a single language like English no single encoding was adequate for all the letters, punctuation, and technical symbols in common use.

Early character encodings also conflicted with one another. That is, two encodings could use the same number for two different characters, or use different numbers for the same character. Any given computer

(especially servers) would need to support many different encodings. Data is passed through different computers or between different encodings, that data runs the risk of corruption. ASCII numbered as from 0 to 127. In extended Uniform Transformation is available that is UTF, UTF-8, UTF-16, UTF-32. It describes how multipattern search can handle the different text encodings encountered in digital forensics and a manner of issues pertaining to proper handling of Unicode in search patterns, The Unicode standard means to be

Universal: Unicode must contain all characters likely to be used in general text interchange.
Efficient: It should synchronize all system for sorting, displaying, searching and editing a text.
Unambiguous: A Unicode code point always contains same character.
Unicode is a multiplatform way, as font developers can use it as reference for their fonts.

II. LITERATURE SURVEY

Support of Unicode by all modern technologies extends application life and broadens integration possibilities for example: Applications supporting Unicode may take advantage of new technologies and integrate with other applications. Widespread industry support provides platform and vendor independence, for example: Microsoft, HP, IBM, Sun, Unisys operating systems, Oracle, Microsoft, Progress databases, and many others support Unicode. Practical design due to the diverse, international, industry and academic membership of the Unicode Consortium for example: Members include computer corporations, software producers, database vendors, research institutions, international agencies, user groups, and linguistic specialists. Easy of conversion from known code pages for example: Unicode is Easy comprehensive character set is superset of existing code pages. Internet-ready for use in E-business for example: Internet standards, such as XML, Perl, Java and JavaScript are Unicode-based.

Online Tools form Unicode consortium

1. Character Classification Tool.
2. Encode decode tool.
3. Bulk_extractor tool.
4. Unicode Character Map.
5. Unicode Input Tool/Converter Firefox Extension.

Unicode Algorithms

B: **Bi-directional text**: left to right, right lo left character read.
C: **Unicode collation algorithm**: customized method to compare two strings.
E: **Unicode equivalence**: standard to allow compatibility with preexisting standard character sets.

I: **ISO 14651**: International string ordering and comparison.

L: **Line wrap and word wrap**: process of breaking a section of text into lines.

Unicode's	Many encoding techniques
ASCII	English. Most widely used before year 2000.
UTF-8	Unicode used in Linux by default, and mostly data on Internet
UTF -16	Unicode is used by Microsoft Windows and Mac, Java programing
GB 18030	Used in China, contains all Unicode chars.
EUC	Extended Unix Code. Used in Japan.
IEC	Series used for most European languages

Table 1. Encoding Techniques for Unicode format. Source: Unicode 11.0

Convert plain text such as letters, sometimes numbers, sometimes and punctuation to different characters from Unicode. Unicode is a character encoding standard that has widespread acceptance. Microsoft software uses Unicode at its core. Unicode just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding could contain enough characters. If your document calls for U+0289 (ʉ) it will be clear to any computer program what the character should be. ASCII defines 128 characters, which map to the numbers 0–127. Unicode defines (less than) 2^{21} characters, which, similarly, map to numbers 0–2^{21}.

```
h1 {
        font-family: Times , "Myownfont";
        font-size: 24px;
}
P {
        font-family: Myownfont; /* If the user is on a any operating system*/
        font-style: normal;
 }
```

III. EXISTING SYSTEM

Unicode already exists with an ASCII 8-bit character at the same some of the unique character sets around 34 thousand are there. For example, this is problem statement because of Search engine optimization they do some tactics techniques to visit their website for commercial use a common people don't know whether this website is authentic or not. Due to that is doing search engine optimization for their website the number of visitors will increase day to day he may give some ads in the website for commercial use. That is the range from 0 to 127 i.e. is each character is 8 bits A is 65, a is 96 is encoded at this area most of the keywords are common in all over websites.

Some Unicode representation for letters.

A is U+0041

a is U+0061

1 is U+0031

is U+0023

Figure 1. Variuos glyphs representaion of lower case a Source: https://en.wikipedia.org/wiki/Glyph

TEXT	ASCII CODE	BINARY REPRESENTAION	UNICODE TEXT UTF-16
A	65	0100 0001	0000 0000 0100 0001
B	66	0100 0010	0000 0000 0100 0010
C	67	0100 0011	0000 0000 0100 0011
D	68	0100 0100	0000 0000 0100 0100
E	69	0100 0101	0000 0000 0100 0101
F	70	0100 0110	0000 0000 0100 0110
G	71	0100 0110	0000 0000 0100 0110
H	72	0100 1000	0000 0000 0100 1000

Table2: representation ASCII and UNICODE formats.

Some Information how font Unicode is used in webpage.

```
h1{     font-family: TimesNewRoman, "Times New Roman", Times, Baskerville, Georgia, serif; font-size: 24px;
  }
P {     font-family: Arial; /* If the user is on a Windows operating system*/font-style: normal;
  }
```

Code: Arial that includes many international characters from the Unicode Standard.

The Web Open Font Format is a font format for use in web pages.

Allograph:

Text	How it is different with the same code
a	Letter U+0061small "A" rendered with a top hook in most fonts
ɑ	U+0251 small Latin alpha never has a top hook.
ɡ	Open tail U+0261 small script "G" never has a loop tail
g	Loop tail U+0067 small "G" rendered with a looptail in some fonts

Table 3: several meanings Source: https://en.wikipedia.org/wiki/Allography

IV. PROPOSED APPROACH

Create a unique character set is the first task, from this area to support Information security, creating private use area code, safe browsing, safe data mining, matching can be done in a constant amount of time. Based on this the results are more accuracy; reliability will be improved. specific character set for websites like government, banking, Institutes, all the services which are under government creating a private use area by separate character set. at same time security can improved at the ATM centers; while entering the pin number or card content information is encoded to our own generated Unicode character set. A user pressed a key combination for "T", which it encodes as U+0054 own character in design phase that is also called as typography based. Same content is encoded in different binary numbers such as Unicode method. Easily we can process the data my matching exacts binary format. Below is sample method we can separated the content in websites by using *charset and encoding can be seen in HTML standard, this code:*

< meta http-equiv="Content-Type" content="text/html; charset=utf-8">

In Our approach we use our character set which is created than easy to find keyword.

After creating own font or character set, will embed in to the coding part as

```
h1 {
        font-family: Myownfont, "Myownfont";
        font-size: 24px;
}
P {
        font-family: Myownfont; /* If the user is on a any operating system*/
        font-style: normal;
 }
```

The feasibility of the approach the integration of our developed search engine.

V.CONCLUSION

separate characterization, classification, extraction of exact data is possible by the new character set by matching the binary data. Avoiding issues like uncertainly handling, discrepancy, searching in huge amount of data, provides Unicode a unique number for every character to make private usage area for the software development, websites, security side is improved to the separate character set as Unicode.

The processing time of the heavy techniques like encoding decoding will be minimized. To exact private code point can be matched in a constant amount of time. The confusion between computers can be limited as much as possible. Web compatibility: Unicode is becoming the universal code page of the Web. Current Web standards require Unicode and rely on it. Ease of world wide deployment. Getting all techniques into one platform.

REFERENCES

1. Unicode Implantation and the meaning for online publishing, Conference on electronic publishing
2. Unicode and color integration technique for encryption and decryption volume 4 issn no. 0976-5697
3. https://ieeexplore.ieee.org/document/954608/.
4. https://ieeexplore.ieee.org/document/7073574/citations.
5. Investigation into using the Unicode standard for primitives of unified characters, Melbourne, Australia, Deaking University
6. Unicode search of Dirty data, or: How I learned to stop worrying and love Unicode technical standard. By Jon, Stewart and Joel Unkelman
7. https://en.wikipedia.org/wiki/Glyph.
8. http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=utconvertq1.
9. https://www.unicode.org/standard/principles.html.
10. http://ergoemacs.org/emacs/unicode_basics.html.
11. https://stackoverflow.com/questions/223360/how-does-gb18030-differ-from-unicode.
12. https://www.unicode.org/standard/WhatIsUnicode.html.
13. https://whatis.techtarget.com/definition/Unicode.
14. Data mining concepts and technique by jiawei han and micheline kamber second edition.