The Cognitive Cause : Metacharacter Spamdexing Bug

[Striking The Base For Bug Proliferation]

By:-Aditya K Sood
Metaeye Security
http://www.metaeye.org

Dated: 13 January 2007
Class:Web Research

Abstract:

This paper consist of the defined cause for the Meta character spamdexing bug.The bug in my previous paper has been fully described and the way it proliferates.I was undertaking this issue and try to analyse the real cause ie where the search engine actually get error prone in the context in which it is coded and modularised.This bug is considered to be as anomaly in real sense if the working of search engine is concernd.But the class of this specific bug should be understood in clear layout.Lets see where the bug comes to play.

Meta Search Engine Definitive Architecture:

To understand the bug proliferation point we have to first look into the architecture of meta search engine in a detailed manner so that the realm of bug gets very clear.This is done to ensure the better learning across different layouts of search panorma.The architecture layout as:

Web Interfacial Arrangement:

The web interfacial arrangement relates to the web interface that is provided to the user for submitting queries and words that are required to be searched via search engine.More specifically this interface can be general or advanced with a difference in the parameter layouts.The search paramters made the query call very definitive when requested from other search engines or the primary base.The queries can be system specific that is how the system is configured for holding that type of queries.This is just an interface for input structure layout in the search engines.

Cognitive Query Parser And Emitter:

Now the second phase starts.At this point what ever the text for the search is entered is being analysed by the local query analyser.It no doubt analyse the text and parse it directly or with any **constraint function** applied to it.The constraint function is somewhat kind of filtering function that filters the text in the prescribed format that is being undertaken by the engine ie in the language relative to search engine.As soon as the text is analysed and parsed it is further emitted by the emitter to the query builder.

Search Engine Query Builder:

The search engine query build the query that is to be provided to remote search engines for performing meta search.This phase designs the query according to specific search engine which is defined on the base of direct or meta specificty.Then on the straight line the query is parsed to remote engines via HTTP header fields.The query is HTTP protocol driven ie the protocol is HTTP for network communication for intra search and inter search related queries.The cache also gets surmount in this aspect that the cache holds some specific results and displayed directly when called again in the same perspective as before.

HTTP Response Splitter:

As the query is passed to remote engines with in a passage of time the response is extracted from that search engine.If this taken in network layout its a client server approach ie a request is submitted ofcourse the response is on your way.So the same network communication model is carried out in the search engine working.Thats why the response is splitted in accordance to the splitting function applied there.So it become most of code oriented where the crucial results are concerned.So response splitter is crucial in the context in which it works.

Search Result Integration:

After the response splitter function is applied the search results are integrated in a centralised layout from where the results are parsed in the back splitter function.This phase is termed to be reverse parsing because the text has to be formatted in the language which is interface oriented ie readable to the user.Again a formatting module is designed for this because the formatting of result is very important for defined layouts.This holds importance in its own context.

Merging And Ranking Standardisation:

The last point of the working of search engine is the merging and ranking of the results recieved.This is crucial only for the result layout and merge the results in very specified way. No doubt this is the last phase of the searching of the text feeded by the user.

So at whole we looked at the basic architecture of the search engine.The prime aim of this to get acquainted with the genericfunctioning of search engine of any kind.This architectural layout is the prime base for search engine working.Now straight forward i will jump where the bug proliferates.

The Reason Of Bug Proliferation:

Why the search engines prone to spamdexing bug.Lets look in the technical viewpoint.The cognitive cause of this bug starts with:

0x01] Search Text Splitting:-

The initiation of bug starts from the point when rogue text is entered into the search box.Here we are talking about the meta characters.so any of the metacharacter is spamdexed ie feed as long strings.The text is splitted into a desired format which is to be feed into the query parser for further building of the search query.The text here is not formatted as it has to be because the formatting function that is applied are not able to handle some of the text ie error prone text.This is because of the coding flaw in the split function that cause the bug to trigger.This sets the starting phase for the bug proliferation.

0x02] Falsified Constraint Function:-

The constraint function is applied at the parsing level ie this is in general language termed to be as the filtering function.The major filtering fucntion checks the text and regular expressions More specifically the filtering is done to handle regular expression which is embedde with text.No doubt the text when fused with meta characters can be filtered very easily.The constraint function some times is not coded as accordance with the text layout.When the bug comes to play the constraint functions are not able to handle the meta character regular expression and hence the text which is entered is considered to be as default case or in general text thereby displaying result based on that.

The Constraint errors:

A] Z*:-

It will display the search starting with Z and display all

B] **:-

what will the function will interpret.The function if not coded in a clear way it will pick up the * by checking it with regular expression but no serial alphabet is there. so the parsing gets wrong the bug start proliferating.

The Coding layout:

```
$temp=some string;
if($temp=~\/\!\@)
{
        //Favour Condition
}
else
{
        //Error Condition.
}
```

So these type of errors are hard to find and corrected.As a result versatile kind of bugs are on the way.

0x03]Reverse Parsing:-

The reverse parsing is atechnique where the text is re parsed and formatted in the language and format which is understandable to user.This is very specific technique because no reverse checks are performed.I lay stress here for Response Checking with the text that is passed through constraint function.Many of the search engine miss this techniques and displayed the result in a straight forward manner.

The spamdexing bug proliferation is the outcome of basic coding errors that remain in there at the core.

Conclusion:

The most of the bugs that persist is due to the coding flaw if application stature is undertaken.The spamdexing bug too favour this way.The bug considered to be as an anomaly. The bug analysis result in development of more and more technology patterns.

Note:

The research layout is the outcome of the spamdexing bug that i encountered in google search engine.This research solely lay importance to the hidden knowledge.Its all done for education purposes.Learning parameter is the limiting factor of this paper.