International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-1)

Research Article | **January 2015**

Search Engine Optimization by Eliminating Duplicate Links

Ms. Pabitha. C
Assistant Professor, Department of CSE,
Valliammai Engineering College, Kattankulathur, India

Abstract:

Duplicate content generally refers to substantive blocks of content within or across domains that either completely matches other content or is appreciably similar. Mostly, this is not deceptive in origin. Examples of non-malicious duplicate content could include: Discussion forums that can generate both regular and stripped-down pages targeted at mobile devices, Store items shown or linked via multiple distinct URLs, Printer-only versions of web pages. In some cases, content are deliberately duplicated across domains in an attempt to manipulate search engine rankings or win more traffic. Deceptive practices like this can result in a poor user experience, when a visitor sees substantially the same content repeated within a set of search results. To avoid these deceptions several methods are proposed in this paper to improve the efficiency of Search engine result page.

Keywords:- SEO, SERP, Duplicate Links, Search Engine, Search Page Optimization, Information Retrieval

I. INTRODUCTION

Search engine optimization (SEO) is the process of affecting the visibility of a website or a web page in a search engine's "natural" or un-paid ("organic") search results. In general, the earlier (or higher ranked on the search results page), and more frequently a site appears in the search results list, the more visitors it will receive from the search engine's users. SEO may target different kinds of search, including image search, local search, video search, academic search, [1] news search and industry-specific vertical search engines.

As an Internet marketing strategy, SEO considers how search engines work, what people search for, the actual search terms or keywords typed into search engines and which search engines are preferred by their targeted audience. Optimizing a website may involve editing its content, HTML and associated coding to both increase its relevance to specific keywords and to remove barriers to the indexing activities of search engines. Promoting a site to increase the number of backlinks, or inbound links, is another SEO tactic.

II. LITERATURE REVIEW

Webmasters and content providers began optimizing sites for search engines in the mid-1990s, as the first search engines were cataloging the early Web. Initially, all webmasters needed to do was to submit the address of a page, or URL, to the various engines which would send a "spider" to "crawl" that page, extract links to other pages from it, and return information found on the page to be indexed. [2] The process involves a search engine spider downloading a page and storing it on the search engine's own server, where a second program, known as an indexer, extracts various information about the page, such as the words it contains and where these are located, as well as any weight for specific words, and all links the page contains, which are then placed into a scheduler for crawling at a later date.

Site owners started to recognize the value of having their sites highly ranked and visible in search engine results, creating an opportunity for both white hat and black hat SEO practitioners. According to industry analyst Danny Sullivan, the phrase "search engine optimization" probably came into use in 1997. [3] On May 2, 2007, [4] Jason Gambert attempted to trademark the term SEO by convincing the Trademark Office in Arizona [5] that SEO is a "process" involving manipulation of keywords, and not a "marketing service." The reviewing attorney basically bought his incoherent argument that while "SEO" can't be trademarked when it refers to a generic process of manipulated keywords, it can be a service mark for providing "marketing services...in the field of computers." [6

Early versions of search algorithms relied on webmaster-provided information such as the keyword meta tag, or index files in engines like ALIWEB. Meta tags provide a guide to each page's content. Using meta data to index pages was found to be less than reliable, however, because the webmaster's choice of keywords in the meta tag could potentially be an inaccurate representation of the site's actual content. Inaccurate, incomplete, and inconsistent data in meta tags could and did cause pages to rank for irrelevant searches. [7] Web content providers also manipulated a number of attributes within the HTML source of a page in an attempt to rank well in search engines. [8]

By relying so much on factors such as keyword density which were exclusively within a webmaster's control, early search engines suffered from abuse and ranking manipulation. To provide better results to their users, search engines had to adapt to ensure their results pages showed the most relevant search results, rather than unrelated pages stuffed with numerous keywords by unscrupulous webmasters. Since the success and popularity of a search engine is determined by its ability to produce the most relevant results to any given search, poor quality or irrelevant search results could lead users to find other search sources. Search engines responded by developing more complex ranking algorithms, taking into account additional factors that were more difficult for webmasters to manipulate. Graduate students at Stanford University, Larry Page and Sergey Brin, developed "Backrub," a search engine that relied on a mathematical algorithm to

rate the prominence of web pages. The number calculated by the algorithm, PageRank, is a function of the quantity and strength of inbound links. [9] PageRank estimates the likelihood that a given page will be reached by a web user who randomly surfs the web, and follows links from one page to another. In effect, this means that some links are stronger than others, as a higher PageRank page is more likely to be reached by the random surfer.

Page and Brin founded Google in 1998. [10] Google attracted a loyal following among the growing number of Internet users, who liked its simple design. [11] Off-page factors (such as PageRank and hyperlink analysis) were considered as well as on-page factors (such as keyword frequency, meta tags, headings, links and site structure) to enable Google to avoid the kind of manipulation seen in search engines that only considered on-page factors for their rankings. Although PageRank was more difficult to game, webmasters had already developed link building tools and schemes to influence the Inktomi search engine, and these methods proved similarly applicable to gaming PageRank. Many sites focused on exchanging, buying, and selling links, often on a massive scale. Some of these schemes, or link farms, involved the creation of thousands of sites for the sole purpose of link spamming. [12]

By 2004, search engines had incorporated a wide range of undisclosed factors in their ranking algorithms to reduce the impact of link manipulation. In June 2007, The New York Times' Saul Hansell stated Google ranks sites using more than 200 different signals. [13] The leading search engines, Google, Bing, and Yahoo, do not disclose the algorithms they use to rank pages. Some SEO practitioners have studied different approaches to search engine optimization, and have shared their personal opinions [14] Patents related to search engines can provide information to better understand search engines. [15]

In 2005, Google began personalizing search results for each user. Depending on their history of previous searches, Google crafted results for logged in users. [16] In 2008, Bruce Clay said that "ranking is dead" because of personalized search. He

opined that it would become meaningless to discuss how a website ranked, because its rank would potentially be different for each user and each search. [17]

In 2007, Google announced a campaign against paid links that transfer PageRank. [18] On June 15, 2009, Google disclosed that they had taken measures to mitigate the effects of PageRank sculpting by use of the nofollow attribute on links. Matt Cutts, a well-known software engineer at Google, announced that Google Bot would no longer treat nofollowed links in the same way, in order to prevent SEO service providers from using nofollow for PageRank sculpting. [19] As a result of this change the usage of nofollow leads to evaporation of pagerank. In order to avoid the above, SEO engineers developed alternative techniques that replace nofollowed tags with obfuscated Javascript and thus permit PageRank sculpting.

Additionally several solutions have been suggested that include the usage of iframes, Flash and Javascript. [20]

In December 2009, Google announced it would be using the web search history of all its users in order to populate search results. [21]

On June 8, 2010 a new web indexing system called Google Caffeine was announced. Designed to allow users to find news results, forum posts and other content much sooner after publishing than before, Google caffeine was a change to the way Google updated its index in order to make things show up quicker on Google than before. According to Carrie Grimes, the software engineer who announced Caffeine for Google, "Caffeine provides 50 percent fresher results for web searches than our last index..." [22]

Google Instant, real-time-search, was introduced in late 2010 in an attempt to make search results more timely and relevant. Historically site administrators have spent months or even years optimizing a website to increase search rankings. With the growth in popularity of social media sites and blogs the leading engines made changes to their algorithms to allow fresh content to rank quickly within the search results. [23]

In February 2011, Google announced the Panda update, which penalizes websites containing content duplicated from other websites and sources. Historically websites have copied content from one another and benefited in search engine rankings by engaging in this practice, however Google implemented a new system which punishes sites whose content is not unique. [24]

In April 2012, Google launched the Google Penguin update the goal of which was to penalize websites that used manipulative techniques to improve their rankings on the search engine. [25]

In September 2013, Google released the Google Hummingbird update, an algorithm change designed to improve Google's natural language processing and semantic understanding of web pages.

III. SERP

A **search engine results page** (**SERP**) is the listing of results returned by a search engine in response to a keyword query. The results normally include a list of items with titles, a reference to the full version, and a short description showing where the keywords have matched content within the page. A SERP may refer to a single page of links returned, or to the set of all links returned for a search query.

3.1 Components of SERP

SERPs of major search engines like Google, Yahoo! and Bing., may include different types of listings: contextual, algorithmic or organic search listings, as well as sponsored listings, images, maps, definitions, videos or suggested search refinements. The major search engines visually differentiate specific content types, such as images, news, and blogs. Many content types have specialized SERP templates and visual enhancements on the main search result page. There are basically three main components of SERP: the actual search query, organic SERP listings, and paid SERP listings.

3.2 Actual Search Query

Also known as 'User Search String', this is the word or set of words that are typed by the user in the search bar of the search engine. Most of the time this search string is a keyword.

3.3 Organic SERP listings

Organic SERP listings are the natural listings generated by search engines based on a series of metrics that determines their relevance to the searched term. Webpages that score well on a search engine's algorithmic test show in this list. These algorithms are generally based upon factors such as the content of a webpage, the trustworthiness of the website, and external factors such as backlinks, social media, news, advertising, etc.

3.4 Paid SERP listings

Paid SERP listings are advertisements, or *sponsored links*, included by search engines in their search results. Websites pay search engines to have their web pages listed here. However, this service is distinct from pay per click, which refers to advertisements placed on websites.

IV. DUPLICATE CONTENT

Search engine spam is any deceitful attempts to deliberately trick the search engine into returning inappropriate, redundant, or poor-quality search results. Many times this behavior is seen in pages that are exact replicas of other pages which are created to receive better results in the search engine. Many people assume that creating multiple or similar copies of the same page will either increase their chances of getting listed in search engines or help them get multiple listings, due to the presence of more keywords.

In order to make a search more relevant to a user, search engines use a filter that removes the duplicate content pages from the search results, and the spam along with it. Unfortunately, good, hardworking webmasters have fallen prey to the filters imposed by the search engines that remove duplicate content. It is those webmasters who unknowingly spam the search engines, when there are some things they can do to avoid being filtered out. In order for you to truly understand the concepts you can implement to avoid the duplicate content filter, you need to know how this filter works.

First, we must understand that the term "duplicate content penalty" is actually a misnomer. When we refer to penalties in search engine rankings, we are actually talking about points that are deducted from a page in order to come to an overall relevancy score. But in reality, duplicate content pages are not penalized. Rather they are simply filtered, the way you would use a sieve to remove unwanted particles. Sometimes, "good particles" are accidentally filtered out.

Knowing the difference between the filter and the penalty, you can now understand how a search engine determines what duplicate content is. There are basically four types of duplicate content that are filtered out:

1. **Websites with Identical Pages** - These pages are considered duplicate, as well as websites that are identical to another website on the Internet are also considered to be spam. Affiliate sites with the same look and feel which contain identical content, for example, are especially vulnerable to a duplicate content filter. Another example would be a website with doorway pages. Many times, these doorways are skewed versions of landing pages. However, these landing pages are identical to other landing pages. Generally, doorway pages are intended to be used to spam the search engines in order to manipulate search engine results.
2. **Scraped Content** - Scraped content is taking content from a web site and repackaging it to make it look different, but in essence it is nothing more than a duplicate page. With the popularity of blogs on the internet and the syndication of those blogs, scraping is becoming more of a problem for search engines.
3. **E-Commerce Product Descriptions** - Many eCommerce sites out there use the manufacturer's descriptions for the products, which hundreds or thousands of other eCommerce stores in the same competitive markets are using too. This duplicate content, while harder to spot, is still considered spam.
4. **Distribution of Articles** - If you publish an article, and it gets copied and put all over the Internet, this is good, right? Not necessarily for all the sites that feature the same article. This type of duplicate content can be tricky, because even though Yahoo and MSN determine the source of the original article and deems it most relevant in search results, other search engines like Google may not, according to some experts.

So, how does a search engine's duplicate content filter work? Essentially, when a search engine robot crawls a website, it reads the pages, and stores the information in its database. Then, it compares its findings to other information it has in its database. Depending upon a few factors, such as the overall relevancy score of a website, it then determines which are duplicate content, and then filters out the pages or the websites that qualify as spam. Unfortunately, if your pages are not spam, but have enough similar content, they may still be regarded as spam.

There are several things you can do to avoid the duplicate content filter

V. ELIMINATING DUPLICACY

Duplicate content is content that appears on the Internet in more than one place (URL). When there are multiple pieces of identical content on the Internet, it is difficult for search engines to decide which version is more relevant to a given search query. To provide the best search experience, search engines will rarely show multiple duplicate pieces of content and thus, are forced to choose which version is most likely to be the original—or best.

Google tries hard to index and show pages with distinct information. This filtering means, for instance, that if your site has a "regular" and "printer" version of each article, and neither of these is blocked with a no index meta tag, we'll

choose one of them to list. In the rare cases in which Google perceives that duplicate content may be shown with intent to manipulate our rankings and deceive our users, we'll also make appropriate adjustments in the indexing and ranking of the sites involved. As a result, the ranking of the site may suffer, or the site might be removed entirely from the Google index, in which case it will no longer appear in search results.
There are some steps you can take to proactively address duplicate content issues, and ensure that visitors see the content you want them to.

Use 301s: If you've restructured your site, use 301 redirects ("Redirect Permanent") in your .htaccess file to smartly redirect users, Googlebot, and other spiders. (In Apache, you can do this with an .htaccess file; in IIS, you can do this through the administrative console.)

Be consistent: Try to keep your internal linking consistent. For example, don't link tohttp://www.example.com/page/ and http://www.example.com/page andhttp://www.example.com/page/index.htm.

Use top-level domains: To help us serve the most appropriate version of a document, use top-level domains whenever possible to handle country-specific content. We're more likely to know that http://www.example.decontains Germany-focused content, for instance, than http://www.example.com/de orhttp://de.example.com.

Syndicate carefully: If you syndicate your content on other sites, Google will always show the version we think is most appropriate for users in each given search, which may or may not be the version you'd prefer. However, it is helpful to ensure that each site on which your content is syndicated includes a link back to your original article. You can also ask those who use your syndicated material to use the noindex meta tag to prevent search engines from indexing their version of the content.

Use Webmaster Tools to tell us how you prefer your site to be indexed: You can tell Google your preferred domain (for example, http://www.example.com or http://example.com).

Minimize boilerplate repetition: For instance, instead of including lengthy copyright text on the bottom of every page, include a very brief summary and then link to a page with more details. In addition, you can use theParameter Handling tool to specify how you would like Google to treat URL parameters.

Avoid publishing stubs: Users don't like seeing "empty" pages, so avoid placeholders where possible. For example, don't publish pages for which you don't yet have real content. If you do create placeholder pages, use thenoindex metatag to block these pages from being indexed.

Understand your content management system: Make sure you're familiar with how content is displayed on your web site. Blogs, forums, and related systems often show the same content in multiple formats. For example, a blog entry may appear on the home page of a blog, in an archive page, and in a page of other entries with the same label.

Minimize similar content: If you have many pages that are similar, consider expanding each page or consolidating the pages into one. For instance, if you have a travel site with separate pages for two cities, but the same information on both pages, you could either merge the pages into one page about both cities or you could expand each page to contain unique content about each city.

The Three Biggest Issues with Duplicate Content

1. Search engines don't know which version(s) to include/exclude from their indices
2. Search engines don't know whether to direct the link metrics (trust, authority, anchor text, link juice, etc.) to one page, or keep it separated between multiple versions
3. Search engines don't know which version(s) to rank for query results

When duplicate content is present, site owners suffer rankings and traffic losses, and search engines provide less relevant results.

SEO Best Practice

Whenever content on a site can be found at multiple URLs, it should be canonicalized for search engines. This can be accomplished using a 301 redirect to the correct URL using the rel=canonical tag (see below) or, in some cases, using the Parameter Handling tool in Google Webmaster Central.

301 Redirect

In many cases the best way to combat duplicate content is to set up a 301 redirect from the "duplicate" page to the original content page. When multiple pages with the potential to rank well are combined into a single page, they not only no longer compete with one another, but also create a stronger relevancy and popularity signal overall. This will positively impact their ability to rank well in the search engines.

Rel="canonical"

Another option for dealing with duplicate content is to utilize the rel=canonical tag. The rel=canonical tag passes the same amount of link juice (ranking power) as a 301 redirect, and often takes up much less development time to implement. The tag is part of the HTML head of a web page. This meta tag isn't new, but like nofollow, simply uses a new rel parameter. For example:

<link href="http://www.example.com/canonical-version-of-page/" rel="canonical" />

This tag tells Bing and Google that the given page should be treated as though it were a copy of the URL www.example.com/canonical-version-of-page/ and that all of the links and content metrics the engines apply should actually be credited toward the provided URL.

The following examples show capitalization errors that cause duplicate content:

1. http://www.simplyhired.com/a/jobs/list/q-**s**oftware+**d**eveloper
2. http://www.simplyhired.com/a/jobs/list/q-**S**oftware+**d**eveloper
3. http://www.simplyhired.com/a/jobs/list/q-**s**oftware+**D**eveloper

The only differences between these URLs are with the capitalization of the words "software" and "developer." A search engine would see all three of these URLs as different pages and would treat them as duplicate content. By implementing the rel="canonical" tool on the 2nd and 3rd instances pointing back to the 1st URL, the search engines would know to treat all of those pages as if they were URL #1.

No index, follow

The meta robots tag with the values "noindex, follow" can be implemented on pages that shouldn't be included in a search engine's index. This allows the search engine bots to crawl the links on the specified page, but keeps them from including them in their index. This works particularly well with pagination issues.

Parameter Handling in Google Webmaster Tools

Google Webmaster Tools allows you to set the preferred domain of your site and handle various URL parameters differently. The main drawback to these methods is that they only work for Google. Any change you make here will not affect Bing or any other search engine's settings.

Set Preferred Domain

This should be set for all sites. It is a simple way to tell Google whether a given site should be shown with or without a www in the search engine result pages.

Additional Methods for Removing Duplicate Content

1. Maintain consistency when linking internally throughout a website. For example, if a webmaster determines that thecanonical version of a domain is www.example.com/, then all internal links should go to http://www.example.com/example.html rather than http://example.com/page.html (notice the absence of www).
2. When syndicating content, make sure the syndicating website adds a link back to the original content. See Dealing With Duplicate Content for more information.
3. Minimize similar content. For example, rather than having one page about raincoats for boys and another page for raincoats for girls that shares 95% of the same content, consider expanding those pages to include distinct, relevant content for each URL. Alternatively, a webmaster could combine the two pages into a single page that is highly relevant for childrens' raincoats.
4. Remove duplicate content from search engines' indices by noindex-ing with meta robots or through removal via Webmaster Tools (Google and Bing).

Rel=Canonical Code Sample

<head> <link rel="canonical" href="http://moz.com/blog/" /> </head>

Meta Robots Code Sample

<head> <meta name="robots" content="noindex, follow" /> </head>

PageRank on the wrong pages can lead to PageRank losses. If you have broken links, dangling links or lots of duplicate content in your website it is certain that you lose valuable link juice. By improving your link structure, by removing deadlinks and dangling links (such as doc, ppt files etc) and by reducing the duplicate content you can increase the number of indexed pages, improve search engine **crawling**, **reduce indexation time** and achieve **better rankings** on the internal pages.

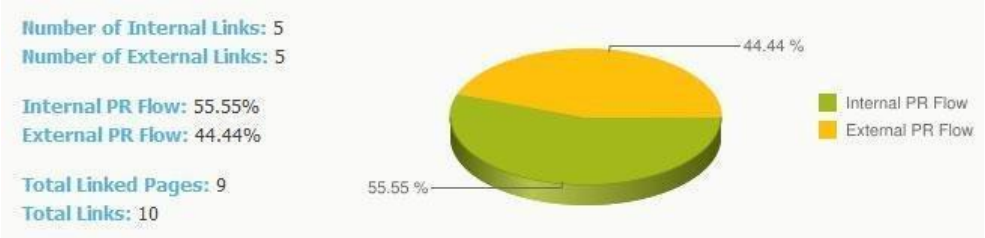

Fig 5.1 Page Duplication due to external/internal links

A great tool that can help you optimize your internal link architecture is the Link Structure Tool. By using it you can **analyze** and **optimize** the link structure of your website, detect and **eliminate** the broken links, optimize the text that is used in the internal anchors, **monitor PageRank flow** and get valuable information on how to reorganize your internal link architecture.

VI. CONCLUSION

The duplicate content filter is sometimes hard on sites that don't intend to spam the search engines. But it is ultimately our job to help the search engines determine that our site is as unique as possible. The paper guides the user in avoiding duplicate contents to a very great extent. This results in optimization in search engine. Thereby it improves the performance.

REFERENCES

[1] Beel, Jöran and Gipp, Bela and Wilde, Erik (2010). "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co.". Journal of Scholarly Publishing. pp. 176–190. Retrieved April 18, 2010.
[2] Brian Pinkerton. "Finding What People Want: Experiences with the WebCrawler"(PDF). The Second International WWW Conference Chicago, USA, October 17–20, 1994. Retrieved May 7, 2007.
[3] Danny Sullivan (June 14, 2004). "Who Invented the Term "Search Engine Optimization"?". Search Engine Watch. Retrieved May 14, 2007. See Google groups thread.
[4] "Trademark/Service Mark Application, Principal Register". Retrieved 30 May 2014.
[5] "Trade Name Certification". State of Arizona.
[6] Bird, Sarah (2008-04-07). "Pulling a Fast One: A Clever Internet Marketer Is Trying to Trademark "SEO"". Retrieved 30 May 2014.
[7] Cory Doctorow (August 26, 2001). "Metacrap: Putting the torch to seven straw-men of the meta-utopia". e-LearningGuru. Archived from the original on April 9, 2007. Retrieved May 8, 2007.
[8] Pringle, G., Allison, L., and Dowe, D. (April 1998). "What is a tall poppy among web pages?". Proc. 7th Int. World Wide Web Conference. Retrieved May 8, 2007.
[9] Brin, Sergey and Page, Larry (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". Proceedings of the seventh international conference on World Wide Web. pp. 107–117. Retrieved May 8, 2007.
[10] "Google's co-founders may not have the name recognition of say, Bill Gates, but give them time: Google hasn't been around nearly as long as Microsoft.". 2008-10-15.
[11] Thompson, Bill (December 19, 2003). "Is Google good for you?". BBC News. Retrieved May 16, 2007.
[12] Zoltan Gyongyi and Hector Garcia-Molina (2005). "Link Spam Alliances" (PDF). Proceedings of the 31st VLDB Conference, Trondheim, Norway. Retrieved May 9, 2007.
[13] Hansell, Saul (June 3, 2007). "Google Keeps Tweaking Its Search Engine". New York Times. Retrieved June 6, 2007.
[14] Danny Sullivan (September 29, 2005). "Rundown On Search Ranking Factors".Search Engine Watch. Retrieved May 8, 2007.
[15] Christine Churchill (November 23, 2005). "Understanding Search Engine Patents".Search Engine Watch. Retrieved May 8, 2007.
[16] "Google Personalized Search Leaves Google Labs". *searchenginewatch.com*. Search Engine Watch. Retrieved September 5, 2009.
[17] "Will Personal Search Turn SEO On Its Ear? | WebProNews". www.webpronews.com. Retrieved September 5, http://en.wikipedia.org/wiki/Wikipedia:No_original_research - Primary.2C_secondary_and_tertiary_sources 2009. [*non-primary source needed*]
[18] "8 Things We Learned About Google PageRank". www.searchenginejournal.com. Retrieved August 17, 2009.
[19] "PageRank sculpting". Matt Cutts. Retrieved January 12, 2010.
[20] "Google Loses "Backwards Compatibility" On Paid Link Blocking & PageRank Sculpting". searchengineland.com. Retrieved August 17, 2009.
[21] "Personalized Search for everyone". Google. Retrieved December 14, 2009.
[22] "Our new search index: Caffeine". Google: Official Blog. Retrieved May 10, 2014.
[23] "Relevance Meets Real Time Web". Google Blog.
[24] "Google Search Quality Updates". Google Blog.
[25] "What You Need to Know About Google's Penguin Update". Inc.com.

AUTHORS

The Author C.PABITHA, was born in Chennai , Kancheepuram District in the year 1986. She completed her M.E Degree in Computer Science & Engineering at Jerusalem College of Engineering, Anna University, Chennai in the year 2009. She got her B.E in Computer Science and Engineering from Valliammai Engineering Colllege, Anna University near Kattankulathur, Kancheepuram District in the year 2007. She is an active member in IEEE Madras Section , ISTE since 2010 and CSI. Her interest in Data Mining has made her publish four papers in National / International Conference and a Journal in Information Retrieval at CiiT International Journal. She has 6 years of teaching experience in academic institutions and she is currently working as a Assistant Professor in Computer Science & Engineering in Valliammai Engineering College, Kattankulathur.

