# Chapter 1: Introduction

## Concept of Search Engines:

The Good news about the Internet and its most visible component, the World Wide Web, is that there are hundreds of millions of pages available, waiting to present information on an amazing variety of topics. The bad news about the Internet is that there are hundreds of millions of pages available, most of them titled according to the whim of their author, almost all of them sitting on servers with cryptic names.

When you need to know about a particular subject, how do you know which pages to read? If you're like most people, you visit an Internet search engine.

Internet search engines are special sites on the Web that are designed to help people find information stored on other sites. There are differences in the ways various search engines work, but they all perform three basic tasks:

* They search the Internet -- or select pieces of the Internet -- based on important words.
* They keep an index of the words they find, and where they find them.
* They allow users to look for words or combinations of words found in that index.

Early search engines held an index of a few hundred thousand pages and documents, and received maybe one or two thousand inquiries each day. Today, a top search engine will index hundreds of millions of pages, and respond to tens of millions of queries per day. In this Project, I'll tell you how these major tasks are performed, and how Internet search engines put the pieces together in order to let you find the information you need on the Web.

## 1.2 What is Search Engine Crawler?

A search engine crawler is a program or automated script that browses the World Wide Web in a methodical manner in order to provide up to date data to the particular search engine. While search engine crawlers go by many different names, such as web spiders and automatic indexers, the job of the search engine crawler is still the same.

The process of web crawling involves a set of website URLs that need to be visited, called seeds, and then the search engine crawler visits each web page and identifies all the hyperlinks on the page, adding them to the list of places to crawl. URLs from this list are re-visited occasionally according to the policies in place for the search engine.

The policies of the search engine can be different for each search engine, and may be a cautionary action to ensure that some of the pages that have been added to the index before have not become spam.

Search engine crawlers have a hard times crawling the web on occasion because the Internet has three main characteristics that make it harder to continually keep the index up to date. Because of the large volume of web pages on the Internet, the fast pace and frequency of change to the pages, and the addition of dynamic pages, many search engine crawlers have a hard time crawling.

These variations produce a massive amount of URLs to crawl, and cause the search engine crawler to prioritize certain web pages and hyperlinks. This prioritization can be summed up in four different search engine crawler policies that are found commonly within all search engines, though they might differ slightly.

* The selection policy is the policy that states which pages to download for the crawling.
* The re-visit policy type is a policy that indicates to a search engine crawler when to check web pages for changes
* The politeness policies are used to inform crawlers as to how to avoid overloading websites to check the URLs
* The parallelization policy is a policy which states how to coordinate distributed web crawlers.

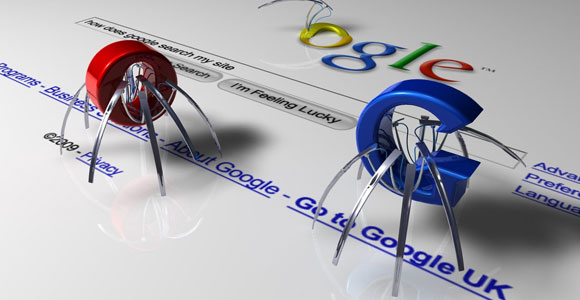


Figure 1: Search Engine Crawlers (GoogleBots)

**Search engine crawlers** generally not only have a good crawling strategy with the policies that allow it to narrow down and prioritize the web pages that need to be crawled, but also need to have a highly optimized architecture. This architecture is used build high-performance systems for search engines that are capable of downloading hundreds of millions of pages over several weeks. This architecture can be followed easily, but must also be ready for high performance results. In a well formed search engine crawler, the web page is taken from the World Wide Web and put through a multi-threaded downloader.

The URLs from this multi-threaded downloader head into a queue, and then pass through a scheduler to prioritize the URLs, finally going through the multi-threaded downloader again so that the text and Meta data ends up in storage.

There are many different professional search engine crawlers available today, such as the Google Crawler, and are used to list the URLs for use in the search engine. Without search engine crawlers, there would be no results for search engine results pages, and new pages would never be listed.

What is robots.txt?

Robots.txt is a text file webmasters create to instruct web robots (typically search engine robots) how to crawl pages on their website. The robots.txt file is part of the the robots exclusion protocol (REP), a group of web standards that regulate how robots crawl the web, access and index content, and serve that content up to users. The REP also includes directives like meta robots, as well as page-, subdirectory-, or site-wide instructions for how search engines should treat links (such as “follow” or “nofollow”).

In practice, robots.txt files indicate whether certain user agents (web-crawling software) can or cannot crawl parts of a website. These crawl instructions are specified by “disallowing” or “allowing” the behavior of certain (or all) user agents.

## 1.3 Don’t Let the Spiders and Web Crawlers Scare You Away!

Does the thought of spiders and web crawlers give you the heebie-jeebies? At media, we make it our job to understand the ins and outs of spiders and web crawlers, but don’t worry, we’re not talking about actual bugs. We’re going to explain the importance of spiders and web crawlers in website optimization, and how they make it possible for people to find the information they need on search engines.

We recently discussed the process of search engine optimization, which is a term that describes procedures used to make a website rank well in search engines. Website optimization involves fixing components of a website to ensure that it is “crawl able” by search engine spiders to improve the overall chances that search engines will index the site.

A search engine spider (also known as a web crawler, web robot, bot, crawler or automatic indexer) is just a program that crawls through links on the internet. They grab content from websites one page at a time and add it to search engine indexes. Spiders follow links from one page to another, and from one website to another which is why links to your site, known as inbound links or backlinks, are so important. Links to your site from other websites will give the search engine spiders more “food” to chew on. The more often they find links to your site, the more likely they are to stop by and visit. Google, especially, relies on spiders to create their large index of web page listings.

## 1.4 How Do Search Engine Spiders Choose Which Webpages to Visit?

Web crawlers comb through websites by looking at the keywords contained on the page, the relevancy of the content and all of the links on the page. They then return that information to the search engine’s server for indexing. Web programmers can also instruct a spider to ignore, and therefore not index, a particular page. When a user enters a search query, the search engine will scour its index, and return the most relevant web pages based on:

* Content: Is your website content relevant with keyword optimized copy, titles, and descriptions?
* Authority: Does your site have valuable content to link to, or do any authoritative websites use your content as a reference/citation?
* Performance: Is your website optimized for fast loading time?

Since search engine spiders find web pages by following links from other pages, the best practice is to earn links from other sites, which will expose your content to search engines naturally. A common misconception is that you must use a submission service to index your website. This used to be the case; however, spammy submissions became problematic, and now are virtually pointless to the practice of modern SEO.

## 1.5 What Search Engine Spiders are NOT Looking for:

If you’re performing dangerous techniques or trying to trick the search engines, red flags are triggered, and you’ll hurt the rankings and visibility of your site. A website can be penalized, and rankings will be negatively affected if search engine algorithms detect these techniques. Items that search engines don’t approve of are:

* Keyword Stuffing: Repeating keywords or trying to hide keywords on your pages.
* Purchasing Inbound Links: Be warned- buying links won’t get you anywhere.
* Bad User Experience: Allowing people to find the content they’re looking for will keep your bounce rate low. Make it as easy as possible for the user to navigate your site.

New websites become available every day, constantly changing the information that is available. Spiders allow search engines and other users to be ensured that their databases are up-to-date. Web crawling is an important method of keeping up with the continuously growing internet.

1.6 Factors for Better Search Engine Optimization.

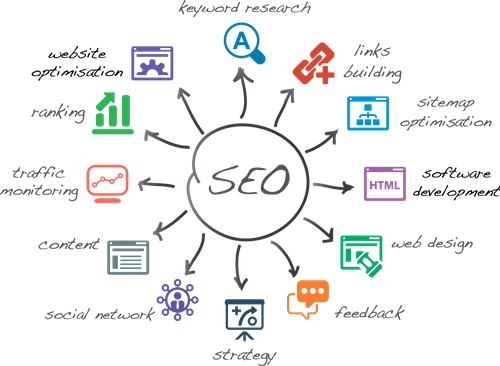


Figure 2: Factors for better Search Engine Optimization

### 1.6.1 On-Page factors:

The way your page is optimized has the most profound effect on its rankings. Here are the page optimization factors that can affect its search visibility:

* Keyword in the title tag. The title Meta tag is one of the strongest relevancy signals for a search engine. The tag itself is meant to give the accurate description of the pages content. Search engines use it to display the main title of a search result. Including a keyword in it will indicate to search engine what to rank the page for.
* Ideally, the keyword should be placed at the start of the title tag. Pages optimized this way will rank better than those with keyword closer to the title’s tag end.
* Keyword in Meta description tag. The importance of the Meta description tag today is often discussed in SEO circles. It is nonetheless still a relevancy signal. It is also crucial for gaining user clicks from search results pages. Including the keyword in it makes it more relevant to a search engine and a searcher.
* Keyword in H1 tag. H1 tag is yet another relevance factor, serving as a description of the pages content. In spite of an ongoing discussion about its importance, it is still a good practice to include your keyword in a unique H1 tag on a page.
* Using keywords in the pages copy. Up until recently, stuffing your page with keywords was a surefire way to increase its rankings for a particular keyword. That’s not the case anymore. Using the keyword in the copy still sends a relevancy signal of what the content is about. How you place it, however, has changed drastically.
* The length of the content. These days’ searchers want to be educated and won’t satisfy with basic information. Google, therefore, looks for authoritative and informative content to rank first. And its common sense that the longer your content is, the greater the chance that you can cover more aspects of your topic. Don’t be shy of writing long but highly useful copy then.
* Duplicate content. Not all factors can influence your rankings in a positive way. Having similar content across various pages of your site can actually hurt your rankings. Avoid duplicating content and write original copy for each page.
* Canonical tag. Sometimes, however, having two URLs with similar content is unavoidable. One of the ways from preventing this from becoming a duplicate content issue is by using a canonical tag on your site. This tag does one simple thing; it tells Google that one URL is equivalent of another, clearly stating that in spite of two pages having the same content, they are in fact one.
* Image Optimization. It’s not only text that can be optimized on a page but other media too. Images, for instance, can send the search engine relevancy signals through their alt text, caption, and description for example.
* Content Updates. Google algorithm prefers freshly updated content. It does not mean that you have to edit your pages all the time. I believe that for commercial pages, such as product descriptions Google recognizes the fact that they are not as time sensitive as blog posts covering recent events. It is wise however to include some strategy to update certain types of content once every 12 months or so.
* Outbound links. Linking to authoritative pages sends trust signals to the search engine. Think of it this way, the only reason why you would send a user to another site is if you wanted them to learn more of the subject. This can be a huge trust factor for Google. Too many outbound links, however, can significantly diminish the page’s PageRank, hurting its search visibility.

Outbound links can affect your rankings but use them in moderation.

* Internal links. Interlinking pages on your site can pass their strength between them.
* Keyword in URL. Including the keyword in the URL slug (that’s the bit that appears after the “.com/“part of the URL) is said to send another relevancy signal to Google.

### 1.6.2 Site factors:

There are certain site-wide factors that can affect your site's search visibility as well:

* Sitemap. A sitemap helps search engine to index all pages on your site. It is the simplest and most efficient way to tell Google what pages your website includes.
* Domain trust. Trust matters. It’s hard no to think that sites Google trusts should rank higher. But how do you build that trust? Brian from Backlinko has a full list of trust factors here. Needless to say, building trust factors of your domain will certainly pay off.
* Server location. Some SEOs believe that a server's location helps to boost rankings for that particular country or region.
* Mobile optimized site. Only a year ago, 46% of searchers used mobile exclusively to research. I believe this number increased exponentially in the last 12 months. It would be no surprise then that having a mobile optimized site would affect rankings in some way.
* Google Search Console integration. Lastly, having your site verified at Google Webmasters Tools is said to help with your sites indexing. Even if that’s not the case, the tool provides valuable data you can use to optimize your site better.

### 1.6.3 Off Page factors:

When ranking your pages, Google looks at factors outside of your site as well. Here are some of the key ones:

* The number of linking domains. The number of domains linking to you is one of the most important ranking factors.
* The number of linking pages. There might be some links from a particular domain to your site; their number is a ranking factor too. However, it is still better to have more links from individual domains rather than from a single domain.
* Domain Authority of linking page. Not all pages are equal. Links to pages with higher domain authority will be a bigger factor than those on low authority domains. Therefore, you should strive to build links from high domain authority websites.
* Link relevancy. Some SEOs believe that links from pages related to your pages topic carry more relevancy for search engines.
* Authority of linking domain. The authority of a domain may be a ranking factor too. For that reason, a link from low authority page on a high authority site will be worth more that from a lower domain authority one.
* Links from a homepage. Similarly, some SEOs believe that links from a home page of a linking domain carry more strength than those on one of its pages.
* A number of do follow vs. nofollow links. Google officially stated that they don’t count nofollow links (link with rel=nofollow attribute attached). Therefore the number of your do follow links should affect your rankings too.
* The diversity of link types. The types of links you build to your site matters too. Too many links of one type may be a spam indicator and impact your rankings negatively.
* Contextual links. It is said that links within the content of the page are worth more than links in a sidebar for instance.
* Link anchor. Anchor text of a link used to be a strong ranking factor. Today it can be utilized as a web spam indicator, negatively impacting your rankings.

### 1.6.4 Domain factors:

Lastly, your domain can affect your rankings as well. Some of the domain signals aren’t as strong as they used to be, there are few things worth paying attention to:

* Domain registration length. Google considers domains registered for longer than a year as more trustworthy. QUOTE.
* Domain history. You may not be the first person who registered the domain. And if your domain has been penalized in the past, its history might affect its current rankings.
* Country TLD extension. If you try to target a particular local market, it is said that having a domain with a country specific TLD (.pl, .co.uk or .ie for instance) will help to achieve better rankings for that location.

There are millions of websites on this earth. Is everyone satisfied with the crawling and indexing rate?? No!! Most people are left constantly wondering why their articles aren’t getting indexed. When you optimize your website based on these factors, Google will have no choice but to crawl and index your page faster and more accurately.

SEO may target different kinds of search, including image search, video search, academic search, news search, and industry-specific vertical search engines. SEO differs from local search engine optimization in that the latter is focused on optimizing a business' online presence so that its web pages will be displayed by search engines when a user enters a local search for its products or service The leading search engines, such as Google, Bing and Yahoo!, use crawlers to find pages for their algorithmic search results.

Pages that are linked from other search engine indexed pages do not need to be submitted because they are found automatically.

Two major directories, the Yahoo Directory and DMOZ, both require manual submission and human editorial review. Google offers Google Search Console, for which an XML Sitemap feed can be created and submitted for free to ensure that all pages are found, especially pages that are not discoverable by automatically following links in addition to their URL submission console.

Yahoo! formerly operated a paid submission service that guaranteed crawling for a cost per click; this was discontinued in 2009.Search engine crawlers may look at a number of different factors when crawling a site.

Not every page is indexed by the search engines. Distance of pages from the root directory of a site may also be a factor in whether or not pages get crawled.

## 1.7 White hat versus black hat techniques

SEO techniques can be classified into two broad categories: techniques that search engines recommend as part of good design, and those techniques of which search engines do not approve. The search engines attempt to minimize the effect of the latter, among them spamdexing. Industry commentators have classified these methods, and the practitioners who employ them, as either white hat SEO, or black hat SEO.

White hats tend to produce results that last a long time, whereas black hats anticipate that their sites may eventually be banned either temporarily or permanently once the search engines discover what they are doing.

An SEO technique is considered white hat if it conforms to the search engines' guidelines and involves no deception. As the search engine guidelines are not written as a series of rules or commandments, this is an important distinction to note.

White hat SEO is not just about following guidelines, but is about ensuring that the content a search engine indexes and subsequently ranks is the same content a user will see. White hat advice is generally summed up as creating content for users, not for search engines, and then making that content easily accessible to the spiders, rather than attempting to trick the algorithm from its intended purpose. White hat SEO is in many ways similar to web development that promotes accessibility, although the two are not identical.

Black hat SEO attempts to improve rankings in ways that are disapproved of by the search engines, or involve deception. One black hat technique uses text that is hidden, either as text colored similar to the background, in an invisible div, or positioned off screen.

Another method gives a different page depending on whether the page is being requested by a human visitor or a search engine, a technique known as cloaking.

Another category sometimes used is grey hat SEO. This is in between black hat and white hat approaches where the methods employed avoid the site being penalized however do not act in producing the best content for users, rather entirely focused on improving search engine rankings.

Search engines may penalize sites they discover using black hat methods, either by reducing their rankings or eliminating their listings from their databases altogether. Such penalties can be applied either automatically by the search engines' algorithms, or by a manual site review.

One example was the February 2006 Google removal of both BMW Germany and Ricoh Germany for use of deceptive practices. Both companies, however, quickly apologized, fixed the offending pages, and were restored to Google's list.

## 1.8 As a marketing strategy

SEO is not an appropriate strategy for every website, and other Internet marketing strategies can be more effective like paid advertising through pay per click (PPC) campaigns, depending on the site operator's goals. Search engine marketing (SEM), is practice of designing, running, and optimizing search engine ad campaigns. Its difference from SEO is most simply depicted as the difference between paid and unpaid priority ranking in search results.

Its purpose regards prominence more so than relevance; website developers should regard SEM with the utmost importance with consideration to PageRank visibility as most navigate to the primary listings of their search. A successful Internet marketing campaign may also depend upon building high quality web pages to engage and persuade, setting up analytics programs to enable site owners to measure results, and improving a site's conversion rate. In November 2015, Google released a full 160 page version of its Search Quality Rating Guidelines to the public, which now shows a shift in their focus towards "usefulness" and mobile search.

SEO may generate an adequate return on investment. However, search engines are not paid for organic search traffic, their algorithms change, and there are no guarantees of continued referrals. Due to this lack of guarantees and certainty, a business that relies heavily on search engine traffic can suffer major losses if the search engines stop sending visitors. Search engines can change their algorithms, impacting a website's placement, possibly resulting in a serious loss of traffic. According to Google's CEO, Eric Schmidt, in 2010, Google made over 500 algorithm changes – almost 1.5 per day. It is considered wise business practice for website operators to liberate themselves from dependence on search engine traffic.

In addition to accessibility in terms of web crawlers (addressed above), user web accessibility has become increasingly important for SEO.

# Chapter 2: Why Python?

## 2.1 Python in Brief.

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable.

It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

* Python is interpreted: Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
* Python is Interactive: You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
* Python is Object-Oriented: Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
* Python is a Beginner's Language: Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

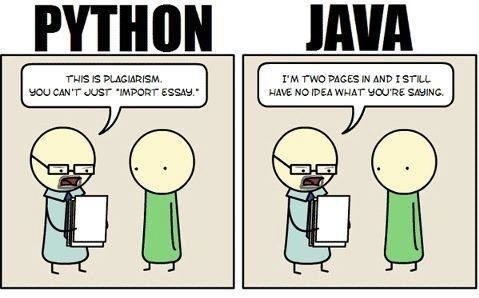


Figure 3: Python vs JAVA

## 2.2 What makes for power in a programming language?

### 2.2.1 Python is Efficient.

A ton of mental energy these days is going into Big Data (both on defining it and on processing it). The more data you have to process, the more important it becomes to manage the memory you use.

Python provides generators, both as expressions and from functions.

Generators allow for iterative processing of things, one item at a time. This doesn’t seem so fancy, until you start to realize that normal iterative processing of a list requires a list. A list takes memory. A really big list takes a lot of memory.

Where this becomes particularly handy is when you have a long chain of processes you need to apply to a set of data. Generators allow you to grab source data one item at a time, and pass each through the full processing chain.

### 2.2.2 Python is fast.

Python is an interpreted language, right? Aren’t they all slow?

The thing is that some amazing work has been done over the past years to improve Python’s performance. My point is that if you want to do high-performance computing, Python is a viable option today.

### 2.2.3 Python is broad.

Python has been around for quite some time now, and is used in nearly every field of endeavor. A reedit thread asked the question how you use Python at work and the answers show tasks ranging from systems automation, testing, and ETL to gaming, CGI and web development.

* Disney uses Python to help power their creative process.
* Mozilla uses Python to explore their extensive code base and releases tons of open source packages built in python.
* Checking Pipit, the public package index for Python, shows 40 thousand add-ons available, across nearly 300 listed topic categories.

Basically, if you want to do something in Python, chances are pretty good someone else already has, and you don’t need to start from scratch.

### 2.2.4 Easy-to-learn:

Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.

### 2.2.5 Easy-to-read:

Python code is more clearly defined and visible to the eyes.

### 2.2.6 Easy-to-maintain:

Python's source code is fairly easy-to-maintain.A broad standard library: Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

### 2.2.7 Interactive Mode:

Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.

### 2.2.8 Portable:

Python can run on a wide variety of hardware platforms and has the same interface on all platforms.

### 2.2.9 Extendable:

You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.

### 2.2.10 Databases:

Python provides interfaces to all major commercial databases.

### 2.2.11 GUI Programming:

Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of UNIX.

### 2.2.12 Scalable:

Python provides a better structure and support for large programs than shell scripting.

# Chapter 3: PROJECT METHODOLOGIES

## 3.1 Introduction.

With the amount of data increasing on the World Wide Web, it becomes extremely important to extract the most relevant information in the shortest span of time. A lot of research is being done to improve the efficiency of search engines by providing crawling algorithms which could traverse through large chunks of data in a short span of time and return the results sorted based on their relevance.

Search engines use algorithms which can sort and rank the results in the order of proximity to the user's query.

* Many algorithms are in use - Breadth First Search, Best First Search, Page Rank algorithm, Genetic algorithm, Naive Bayes classification algorithm to mention a few.

Whatever information we get might not be completely useful. There might be a scenario where the website might contain a large number of popularly searched keywords, just to increase the number of hits on their website.

So there is always the challenging task to get search engines which provides relevancy, robustness and the ability to download optimal number of pages.

With heuristic approach being compared to native techniques of web crawling, we focus on a comparative study between these approaches.

The comparison will be done on simulated web environment and the time to search is the time taken to reach the most relevant page from any initial page.

## 3.2 Literature Survey

When a data is searched, hundreds of thousands of results appear. Users do not have the persistence and stretch to go through each and every page listed.

So search engines have a big job of sorting out the results, in the order of interest to the user within the first page of appearance and a quick summary of theRetrieving effective content from the Web is a crucial task because it heavily influences the perceived effectiveness of a search engine.

Users often look at only a few top hits, making the precision achieved by the ranking algorithm of paramount importance. Early search engines ranked pages principally based on their lexical similarity to the query.

The key strategy was to devise the best weighting algorithm to represent Web pages and query in a vector space, so that closeness in such a space would be correlated with semantic relevance e information provided on a page.Web Crawler is a program/software or automated script which browses the World Wide Web in a methodical, automated manner.

Crawlers have bots that fetch new and recently changed websites, and then indexes them. By this process billions

of websites are crawled and indexed using algorithms (which are usually well-guarded secrets) depending on a number of factors.

Several commercial search engines change the factors often to improve the search engines process.

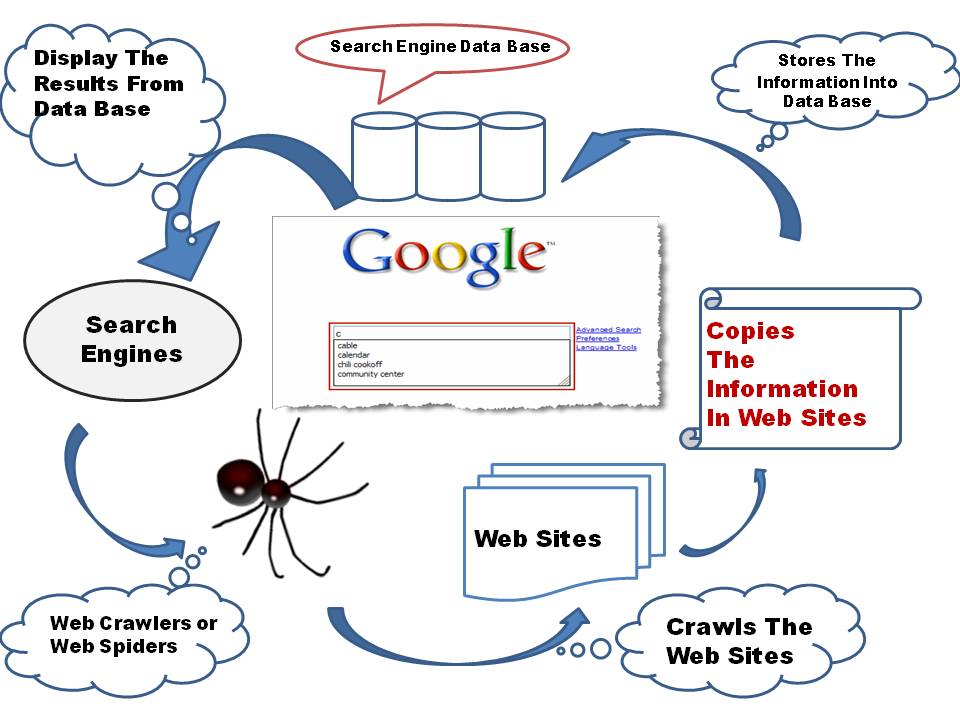


Figure 4: Components of a Web Search System with Web Crawler.

The basic procedure executed by any web crawling algorithm takes a list of seed URLs as its input and repeatedly executes the following steps:

* Remove a URL from the URL list.
* Download the corresponding page.
* Check the Relevancy of the page.
* Extract any links contained in it.
* Add these links back to the URL list. After all URLs are processed, return the most relevant page.

## 3.3 GENERAL CRAWLING STRATEGIES:

There are many highly accomplished techniques in terms of Web crawling strategy. The researcher describe most relevant here:

### 3.3.1Breadth-First Crawling

The Breadth-First search algorithm performs the unique search around the neighbor nodes(hyperlinks). It start by following the root node (Hyperlink) and scans the all the neighbor nodes at the initial level. If the targeted search is achieved then the scanning is stopped otherwise it leads to the next level.Such types of algorithms are best suited where the branches are small and resultant objective is identical. When the branches or tree is very deep then this algorithm will not perform well, i.e. all path traversals leads to the same resultant node.

### 3.3.2 Depth First Crawling

The Depth First search algorithm starts searching the objective from the root node and traverse next to its child node, If there are more than one child node, then left most node is given highest priority and traverse deep until no more child node is present.

Then it starts from the next unvisited node and then continues in a similar manner. By using this algorithm the assurance of scanning of all node is achieved but when the number of child node is large then this algorithm takes more time and might go in to infinite.

### 3.3.3 Targeted Crawling

Search engines use random crawling process in order to target a certain type of page, e.g. pages on a specific topic or in a particular language, images, mp3 files, geo location, and domain specific or scientific papers. In addition to these heuristics, more generic approaches have been suggested. They are based on the analysis of the structures of hypertext links and techniques of learning: the objective here is being to retrieve the greatest number of pages relating to a particular subject by using the minimum bandwidth. Most of the studies cited in this category do not use high performance.

### 3.3.4 Page Rank Algorithm

This algorithm works on the importance of the web pages. It calculates inlinks or backlinks to that page. Then the page rank is given to each page as per bellow formula. PR(A) = (1-d)+d(PR(T1)/C(T1))+…PR(Tn)/C(Tn)) Where, PR(A) :-> Page Rank of Site. d :-> damping factor. T1,….Tn :-> no. of links.

After determining a page rank of a website the index has been generated to show the relevant on a website contain to the search keywords.

## Table 1: Comparison on Crawling Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Search Pattern** | **Benefits** | **Drawbacks** |
| Breadth First Crawling | Scans neighbor node from root level,  if result not achieved then go to next level. | where the branches are small and resultant objective is identical. | When the branches or tree is very deep then goes into infinite. |
| Depth First Crawling | Scans from the root node and traverse next to its child leftmost node | the assurance of scanning of all node is achieved | Takes more time when the child node is large. |
| Targeted Crawling | Uses random (heuristics) crawling process | retrieves the greatest number of pages relating to a particular subject by using the minimum bandwidth. | Takes more time when specific topics are very large. |
| Page Rank Algorithm | works on the importance of the web pages. It calculates inlinks or backlinks to that page | More accurate search result | Difficult to manage and update page index repository |

## 3.4 CRAWLING ALGORITHMS

We now discuss a number of crawling algorithms that are suggested in the literature. Note that many of these algorithms are variations of the best-first scheme.

The difference is in the heuristics they use to score the unvisited URLs with some algorithms adapting and tuning their parameters before or during the crawl.

### 3.4.1 Naive Best-First Crawler

A naive best-first represents the collection of fetched URLs as a vector. In the study of crawler evaluation the Naïve Best First Crawler was one of the most evaluated algorithm by the authors. In this algorithm the cosine similarity of the page with the query or description provided by the user is calculated.

Then weight can be generated for unvisited URLs on each page by this cosine value. After that the URL is inserted to the vector based on the cosine weight. The crawler iterates by picking the best URL in the vector to crawl and returns new unvisited URLs that are again inserted in to the vector based on the cosine weight of the parent page.

### 3.4.2 Shark Search

Shark Search is a more aggressive version of Fish Search with some improvements. In Fish-Search crawling, the crawler search more broadly in the areas of the web where number of relevant page found is more. At the same time the crawler skip the areas where the relevant pages are not found.

SharkSearch uses a similar valued function to measure the relevance as opposite to the binary relevance function of FishSearch.

In addition of these, SharkSearch has a more sophisticated concept of potential scores for the links in the crawl frontier. The Potential score of links is influenced by anchor text, link context, and inherited score from ancestors (incoming and outgoing links URLs of the page).

### 3.4.3 Focused Crawler

Chakrabarti has invented a focused crawler based o a hypertext classifier et al. [9, 6]. The main aim of the crawler is to categorize the crawled pages in to different topic based categories. To begin, the crawler requires a topic classification such as Yahoo. User can also provide their interested search keywords or URLs.

Examples provided by the users get categorized in to different categories of classification. The crawler uses the Bayesian classifier to set the probability of a page that the page will belong to which category in the classification.

Then the crawling process is similar to that of Naïve Best-First Crawling, it picks up the page with a highest match with the user query and starts the crawling for unvisited URLs found from that page and inserts in to the vector for further classification.

### 3.4.4 Context Focused Crawler

Sometimes if we are looking for specific topic i.e. “Computer Architecture”, that word or topic may not be on the home page of computer science website. To reach at the topic we need to go first to the home page of computer science website then move to faculty pages which may lead to specific topic.

To estimate the relevance of the link distance between a crawled page and the specific page the Context Focused Crawler is used, unlike focused crawler, the Context Focused Crawler is so much advance.

Such pages have given low priority in naive best-first crawler and may never crawl it again, this crawler can estimate that the relevant page of “Computer Architecture” is two link away from the “About Computer” page then the home page of computer science website.

And the highest priority can be given to About Computer by using context graph of layers corresponding to seed page.

### 3.4.5 Info Spiders

In Info Spider [21, 23] algorithm, an adaptive population of agents perform the search for pages relevant to the topic given by the search query. By using an adaptive query list and a neural network each agent follows the crawling loop to decide which links to crawl next. The algorithm provides an exclusive frontier for each agent.

While each thread has its own frontier to fetch the pages, the crawler was limited to following the links on the current page and it was outperformed by the naïve best-first crawler on a number of evaluation standards.

As taking inspiration from the naïve best-first algorithm, many improvements have been made in the Info Spider. And the redesigned version has been found to outperform on crawling task that are longer than ten thousand pages.

An agent consists of a list of search keywords and a neural network to evaluate new links. The occurrence of keyword is weighted while traversing each link based on the nearness of the link from the given keyword.

After fetching new page, the agent receives “energy” in proportion to the similarity between its keyword and the new page. A back-propagation algorithm is used to learn prediction of similarity estimation. In this section we have presented a variety of crawling algorithms, most of which are variations of the best-first scheme.

The readers may pursue Menczer et.al. for further details on the algorithmic issues related with some of the crawlers.

## 3.5 DESIGN IMPLEMENTATION

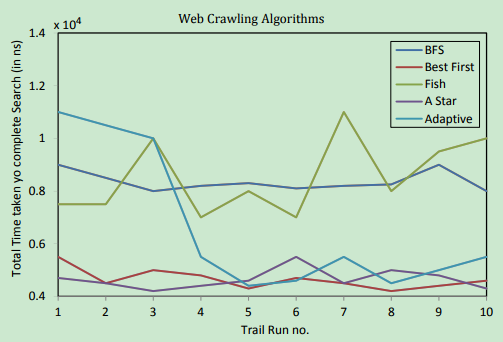
In the implementation of these algorithms, we have made some assumptions related to Relevancy Calculation and Link Extraction. To calculate the relevancy of different documents, random relevancy values between 0 and 1 are generated. Whenever a page is extracted, a random relevancy is generated corresponding to it. 

Figure 5:Search Time of Various Web Crawling Algorithms

All Link extractions are done by traversing a graph and maintaining lists of unvisited pages. Since the Web is a directed graphical structure, for implementation purpose a graph is considered where each page is represented by a node and link from one page to another is represented by an edge.

In the implementation process, web crawler is provided with a seed URL (Graph Node) and then it uses the algorithm to traverse the graph.

The time taken for each traversal is the difference between the start time and end time of each algorithm. The graph above, Fig. 2; shows an approximate trend in the running time of these algorithms.

The actual running time largely depends on the initial node, the relevancy of the nodes and the network structure.

Though the graph may not be completely accurate, it shows the general trend found while running the algorithms with different initial nodes. The X - axis represents the trail number, such as the particular instance in the sequence of trails and Y - axis represents the total time taken in nanoseconds to complete the journey from the initial node to the goal node.

Even though the difference in total time taken to search might look large from the graph, the actual difference is very less as all the calculations are done in nanoseconds.

# Chapter 4: Result (Output)

## 4.1 Indexation of EBAY (<http://www.ebay.in/>)

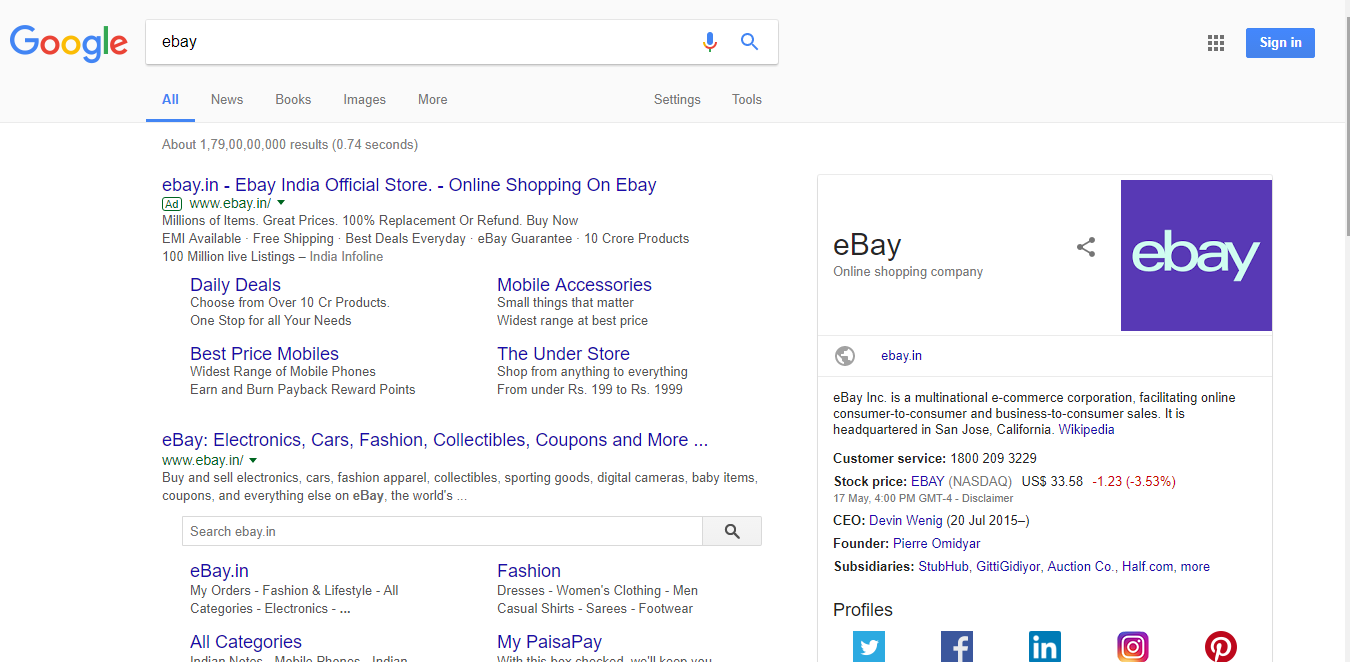


Figure 6: Search Result of EBAY showing Page Title, Meta Description

## 4.2 Robots.txt file has been found.

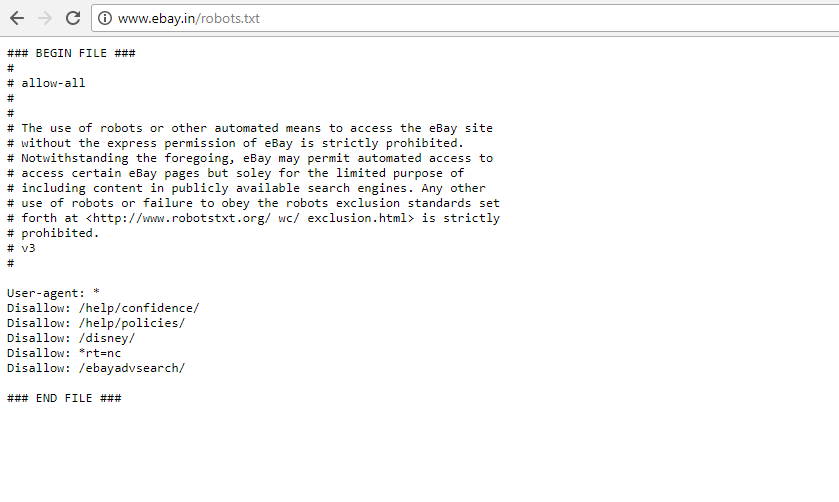


Figure 7: Screenshot of ROBOTS.TXT File has been found.

## 4.3 Sitemaps has been found in HTML not in XML.

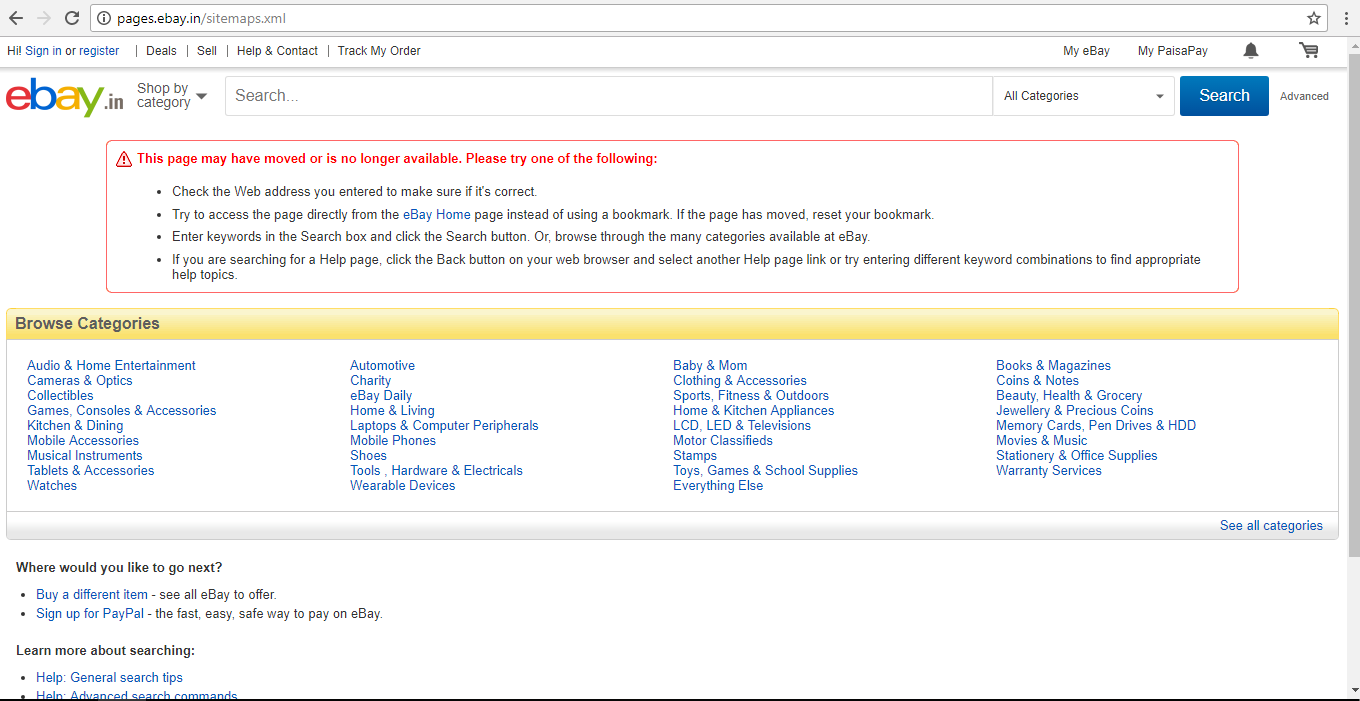


Figure 8: Screenshot for showing that Sitemaps are not found in XML.

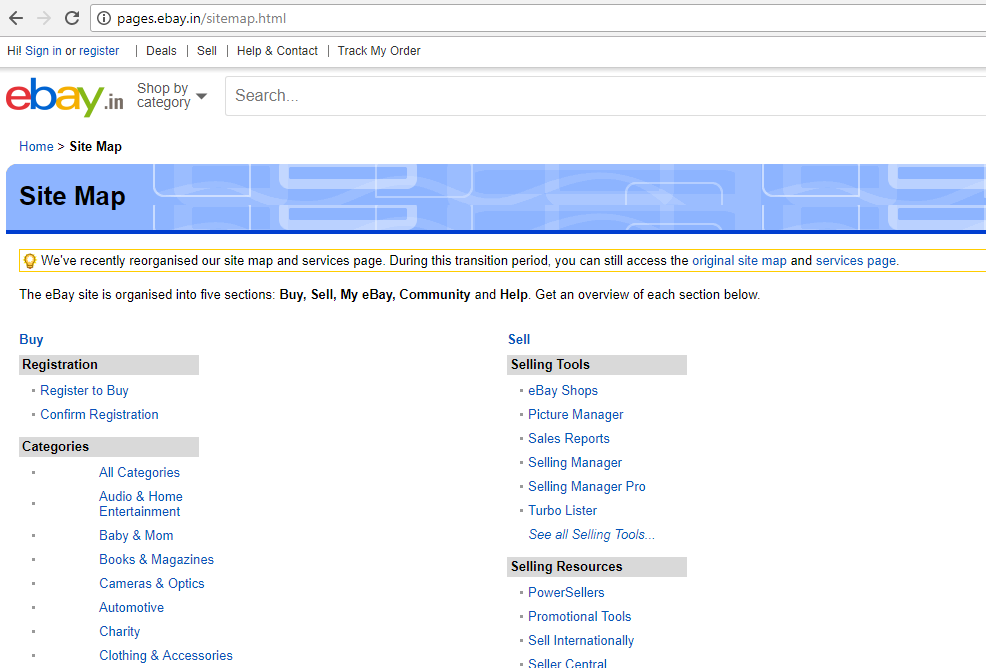


Figure 9:Screenshot showing that Sitemaps are found in .HTML

## 4.4 Good amount of TEXT found on Web Page.

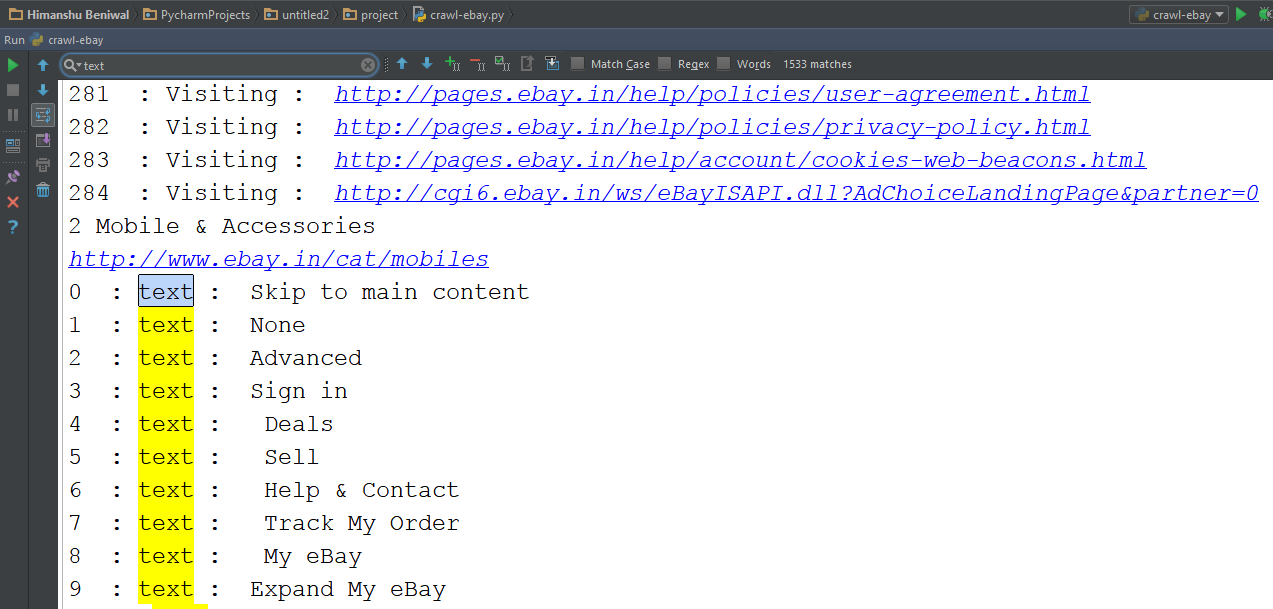


Figure 10. Screenshot showing the text occurred 1533 times while crawling.

## 4.5 Links found while crawling.

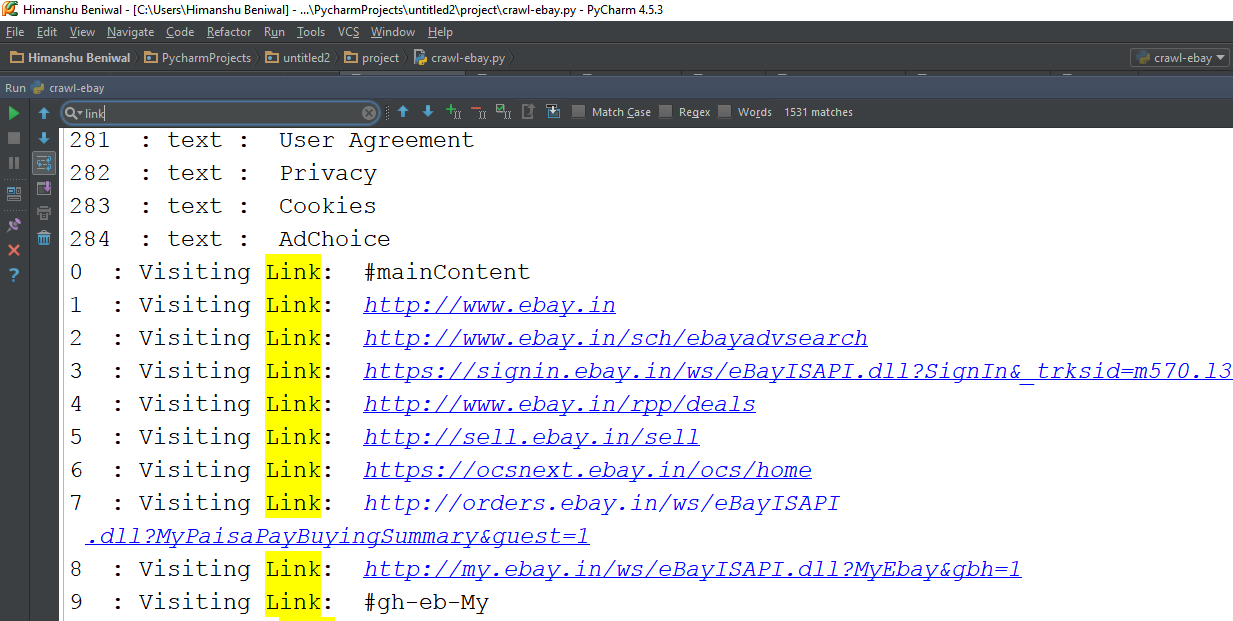


Figure 11: Screenshot showing the 'Link' occurred 1531 times while crawling.

## 4.6 Main Title (specific Class) which were Crawled.

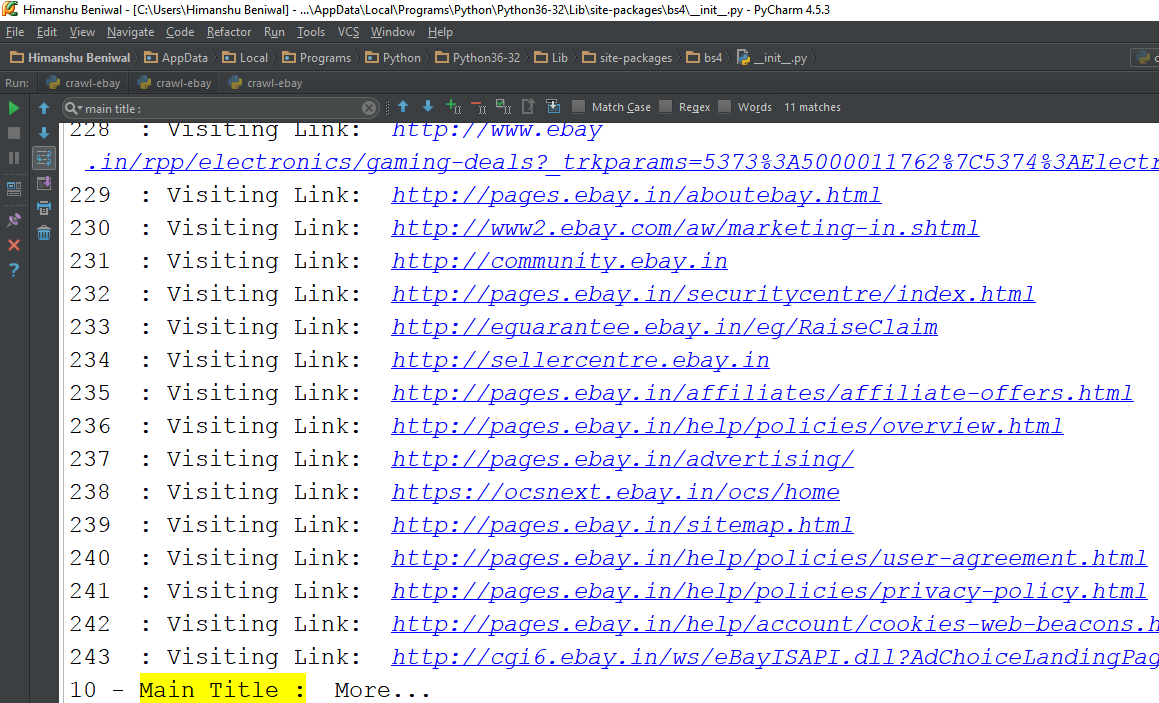


Figure 12: Main Title (specific Class: 'rt') were Crawled.

## 4.7 What a Crawler See VS What a User See?

### 4.7.1 What a Crawler Crawls?

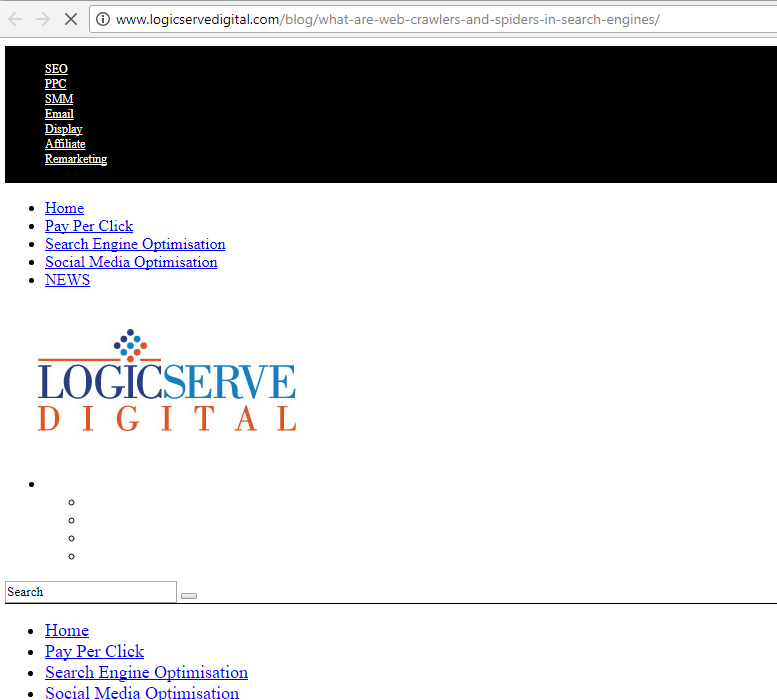


Figure 13: This is What a Crawler see (No JAVASCRIPT or CSS).

### 4.7.2 What a User is Able to See?

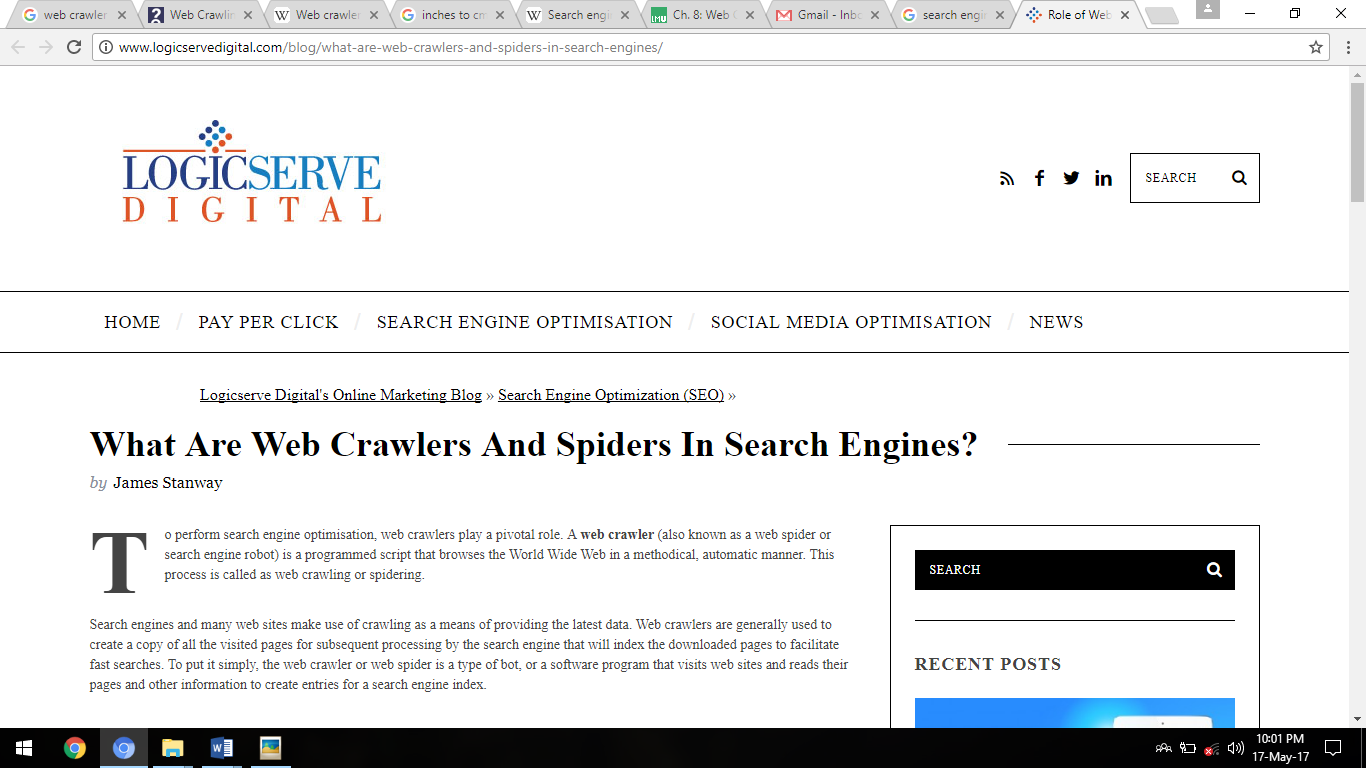


Figure 14: Screenshot for What a User is able to see (including JAVASCIPT or CSS)

## 4.8 Accessibility Overview for Web Page (using Tool).

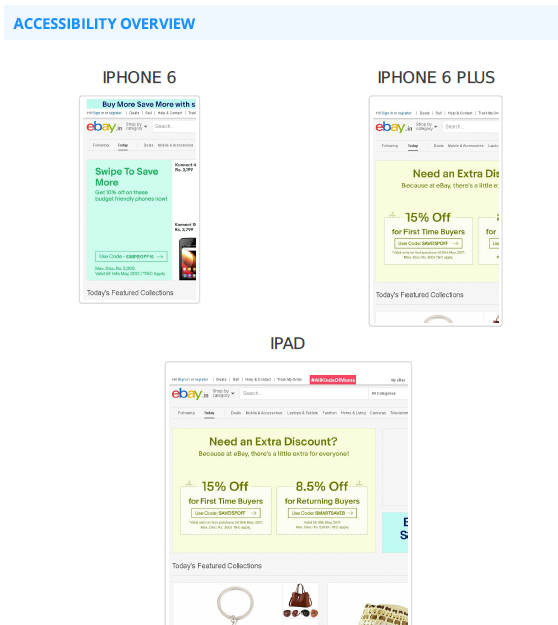


Figure 15: Screenshot for showing how different the page is view in different Devices.

# Chapter 5: Conclusion

A crawler is a program that retrieves and stores pages from the Web, commonly for a Web search engine. A crawler often has to download hundreds of millions of pages in a short period of time and has to constantly monitor and refresh the downloaded pages.

In addition, the crawler should avoid putting too much pressure on the visited Web sites and the crawler's local network, because they are intrinsically shared resources.

## 5.1 Overview of Search Engine Optimization.

Search Engine Optimization – SEO is the art and science of optimizing your web pages to be found by the search engines like Google, Yahoo, and Bing etc. But why part art, part science? The science part of Search Engine Optimization (SEO) deals with technology and all the engineering & technicalities of a website such as code, information architecture & user experience etc.

While Art is all about understanding the business, and the customer behavior and try to positively persuade it by adding value to them (in the form of a product, service or a solution) and getting value in return (as leads, sales or branding).

Search Engine Optimization (SEO) plays a fundamental role in scaling up sales and speed up the growth of a completely online business (such as Flipkart.com or Snapdeal.com) or an offline business (such as KFC).

## 5.2 Reasons to Choose SEO as Your Career:

### 5.2.1 High demand for SEO services

With an increase in demand for SEO services, employment opportunities in the industry are also increasing. Now, when sites have increased in number and doing great business, an expert SEO has become the need of the companies. Therefore, you can expect more job opportunities in the Digital Marketing industry.

### 5.2.2 A LOT of people have made a successful SEO career

There are many living proofs that SEO is a viable business. The list is too long to be quoted here but some of the names include Rob from Blackwood Productions, Jill Wahlen from High Rankings, Rand Fishkin from SEO Moz and many others.

### 5.2.3 Search Engine Optimizers make Good Money

SEO is a profession that can be practiced while working for a company or as a solo practitioner. There are many jobboards like Dice and Craigslist that publish SEO job advertisements. It is worth noting that the compensation for SEO employees is equal to or even higher than that of developers, designers and marketers. Salaries over $80K per annum are not an exception for SEO jobs.

As a solo SEO practitioner you can make even more money. Almost all freelance sites have sections for SEO services and offers for $50 an hour or more are quite common. If you are still not confident that you can work on your own, you can start a SEO job, learn a bit and then start your own company.

If you already feel confident that you know a lot about SEO, you can take this quiz and see how you score. Well, don’t get depressed if you didn’t pass – here is a great checklist that will teach you a lot, even if you are already familiar with SEO.

### 5.2.4 Only Web–Designing MAY NOT be enough.

Many companies offer turn–key solutions that include Web design, Web development AND SEO optimization. In fact, many clients expect that when they hire somebody to make their site, the site will be SEO friendly, so if you are good both as a designer and a SEO expert, you will be a truely valuable professional.

On the other hand, many other companies are dealing with SEO only because they feel that this way they can concentrate their efforts on their major strentgh – SEO, so you can consider this possibility as well.

### 5.2.5 Logical step ahead if you come from marketing or advertising

The Web has changed the way companies do business, so to some extent today’s marketers and advertisers need to have at least some SEO knowledge if they want to be successful. SEO is also a great career for linguists.

### 5.2.6 Lots of Learning

For somebody who comes from design, development or web administration, SEO might look not technical enough and you might feel that you will downgrade if you move to SEO. Don’t worry so much – you can learn a LOT from SEO, so if you are a talented techie, you are not downgrading but you are actually upgrading your skills packages.

### 5.2.7 SEO is already recognized as a career

Finally, if you need some more proof that SEO is a great career, have a look at the available SEO courses and exams for SEO practitioners. Well, they might not be a CISCO certification but still they help to institutionalize the SEO profession.

## 5.3 Ugly face of SEO

### 5.3.1 Dependent on search engines

It is true that in any career there are many things that are outside of your control but for SEO this is a rule number one. Search engines frequently change their algorithms and what is worse – these changes are not made public, so even the greatest SEO gurus admit that they make a lot of educated guesses about how things work. It is very discouraging to make everything perfect and then to learn that due to a change in the algorithm, your sites dropped 100 positions down. But the worst part is that you need to communicate this to clients, who are not satisfied with their sinking ratings.

### 5.3.2 No fixed rules

Probably this will change over time but for now the rule is that there are no rules – or at least not written ones. You can work very hard, follow everything that looks like a rule and still success is not coming. Currently you can’t even rely on bringing a search engine to court because of the damages they have done to your business because search engines are not obliged to rank high sites that have made efforts to get optimized.

### 5.3.3 Rapid changes in rankings

But even if you somehow manage to get to the top for a particular keyword, keeping the position requires constant efforts. Well, many other businesses are like that, so this is hardly a reason to complain – except when an angry customer starts shouting at you that this week their ratings are sinking and of course this is all your fault.

### 5.3.4 SEO requires Patience

The SEO professional and customers both need to understand that SEO takes constant effort and time. It could take months to move ahead in the ratings, or to build tens of links. Additionally, if you stop optimizing for some time, most likely you will experience a considerable drop in ratings. You need lots of motivation and patience not to give up when things are not going your way.

### 5.3.5 Black hat SEO

Black hat SEO is probably one of the biggest concerns for the would–be SEO practitioner. Fraud and unfair competition are present in any industry and those who are good and ethical suffer from this but black hat SEO is still pretty widespread. It is true that search engines penalize black hat practices but still black hat SEO is a major concern for the industry.

## Why SEO industry is growing?

1. Increased use of search engines to look for information on web year on year.

2. Large brand advertisers are using SEO actively in India & Internationally.

3. Traditional or Direct marketers are increasingly adopting SEO.

4. Growth in the number of local search request with mobile web searches.

5. Organic SEO listings on Google search result page is more trustworthy than Ads.

6. Google are fast evolving and making demographic targeted & Geo targeted search results.

## 5.4 SEO Salary & Benefits

In Asian countries like India the average salary for SEO is in between Rs 1.8 lacs – Rs 4 lacs per year for freshers.

Like any other discipline; the salary depends on various different factors like No. of years’ Experience, Employer Type, City, Company Size, Certifications etc.

There are thousands of job openings in the Digital marketing industry and many of the job descriptions include experience in search engine optimization. Need for both entry-level & experienced SEO/SEM professionals will be growing until search engines exist in this world.

## 5.5 Scope of SEO

Online industry is growing everyday & SEO is playing a vital role in bringing the visitors to the website and all the companies are spending a huge sum in SEO. It’s one of the most happening areas and there are a good number of job opportunities in the field of SEO.

Every website is looking to procure SEO services for itself either internally or externally depending on its budgets and expertise at hand.

SEO has very bright future for you if you have great analytic skills, and you are dynamic. You need to keep up with the latest trend in the field, have passion for research and be able to formulate clever and creative marketing strategy to beat your competitor, and have an edge over all others.

## 5.6 SEO Job Categories

Many of the SEO/SEM firms, particularly the larger firms, have employees in these job categories:

• Analytics

• Business Management/Development

• Link Building

• Event Management

• Organic Search Engine Optimization

• Social Media Analyst

• Web Development/Programming

• Web Design

• Offline Marketing/Advertising

• Public Relations/Reputation Management

• Paid Search/PPC Management

• Web Design

• Writing/Blogging.

## References:

1. WIKIPEDIA(<https://en.wikipedia.org/wiki/Search_engine_optimization#As_a_marketing_strategy>)
2. Why Make Career In Search Engine Optimization (SEO) - Scope, Pros, Cons & Salary (<https://www.linkedin.com/pulse/why-make-career-search-engine-optimization-seo-scope-gulshan-sirohi>)
3. Crawl, Index, Rank, Repeat: A Tactical SEO Framework (<https://searchenginewatch.com/sew/news/2064349/crawl-index-rank-repeat-a-tactical-seo-framework>)
4. Web Crawler (<https://en.wikipedia.org/wiki/Web_crawler>)
5. How Stuff Works (<http://computer.howstuffworks.com/internet/basics/search-engine1.htm>).
6. Youtube (<https://www.youtube.com/watch?v=CDXOcvUNBaA>).
7. Wikipedia (<https://en.wikipedia.org/wiki/Search_engine_optimization>).
8. Blog-ShoutMeLoud (<https://www.shoutmeloud.com/google-crawling-and-indexing.html>)
9. Search Engine land – ([www.searchengineland.com/guide/what-is-seo](http://www.searchengineland.com/guide/what-is-seo)).

## Annexure

#Program for Web Crawler

#importing Modules to use :

import requests

from bs4 import BeautifulSoup

deftrade\_spider(max\_pages): #max pages for Different number of pages

page = 1 #intializing the page value to start from 1.

while page <= max\_pages:

url = "http://www.ebay.in" #Enter Link to be Crawl

source\_code = requests.get(url) #Requesting the URl and Saving in Source\_Code

plain\_text = source\_code.text # just get the code, no headers or anything

soup = BeautifulSoup(plain\_text) # BeautifulSoup objects can be sorted through easy

letter=0

for link in soup.findAll('a',{'class':'rt'}):

‘’’Define Class to look on Webpage and Crawling should be on this class ‘’’

href = link.get('href') #getting link

title = link.string # just the text, not the HTML

print(letter,'- Main Title : ',title)

letter=letter+1

print('Link for Main Title to be Crawl : ',href)

get\_single\_item\_data(href) #function to Crawl more Inner Level link

page += 1

defget\_single\_item\_data(item\_url):

source\_code = requests.get(item\_url)

plain\_text = source\_code.text

soup = BeautifulSoup(plain\_text)

# if you want to gather information from that page

count=0

foritem\_name in soup.findAll('a'):

print(count," : text : ",item\_name.string)

count=count+1

# if you want to gather links for a web crawler

count=0

for link in soup.findAll('a'):

href =link.get('href')

print(count," : Visiting Link: ",href)

count=count+1

trade\_spider(1)

Contents

[Chapter 1: Introduction 1](#_Toc482882819)

[1.1 Concept of Search Engines: 1](#_Toc482882820)

[1.2 What is Search Engine Crawler? 1](#_Toc482882821)

[1.3 Don’t Let the Spiders and Web Crawlers Scare You Away! 3](#_Toc482882822)

[1.4 How Do Search Engine Spiders Choose Which Webpages to Visit? 3](#_Toc482882823)

[1.5 What Search Engine Spiders are NOT Looking for: 4](#_Toc482882824)

[1.6.1 On-Page factors: 5](#_Toc482882825)

[1.6.2 Site factors: 6](#_Toc482882826)

[1.6.3 Off Page factors: 6](#_Toc482882827)

[1.6.4 Domain factors: 7](#_Toc482882828)

[1.7 White hat versus black hat techniques 8](#_Toc482882829)

[1.8 As a marketing strategy 9](#_Toc482882830)

[Chapter 2: Why Python? 10](#_Toc482882831)

[2.1 Python in Brief. 10](#_Toc482882832)

[2.2 What makes for power in a programming language? 11](#_Toc482882833)

[2.2.1 Python is Efficient. 11](#_Toc482882834)

[2.2.2 Python is fast. 11](#_Toc482882835)

[2.2.3 Python is broad. 11](#_Toc482882836)

[2.2.4 Easy-to-learn: 11](#_Toc482882837)

[2.2.5 Easy-to-read: 11](#_Toc482882838)

[2.2.6 Easy-to-maintain: 11](#_Toc482882839)

[2.2.7 Interactive Mode: 12](#_Toc482882840)

[2.2.8 Portable: 12](#_Toc482882841)

[2.2.9 Extendable: 12](#_Toc482882842)

[2.2.10 Databases: 12](#_Toc482882843)

[2.2.11 GUI Programming: 12](#_Toc482882844)

[2.2.12 Scalable: 12](#_Toc482882845)

[Chapter 3: PROJECT METHODOLOGIES 13](#_Toc482882846)

[3.1 Introduction. 13](#_Toc482882847)

[3.2 Literature Survey 13](#_Toc482882848)

[3.3 GENERAL CRAWLING STRATEGIES: 14](#_Toc482882849)

[3.3.1Breadth-First Crawling 15](#_Toc482882850)

[3.3.2 Depth First Crawling 15](#_Toc482882851)

[3.3.3 Targeted Crawling 15](#_Toc482882852)

[3.3.4 Page Rank Algorithm 15](#_Toc482882853)

[Table 1: Comparison on Crawling Algorithms 15](#_Toc482882854)

[**Algorithm** 15](#_Toc482882855)

[**Search Pattern** 15](#_Toc482882856)

[**Benefits** 15](#_Toc482882857)

[**Drawbacks** 15](#_Toc482882858)

[3.4 CRAWLING ALGORITHMS 16](#_Toc482882859)

[3.4.1 Naive Best-First Crawler 16](#_Toc482882860)

[3.4.2 Shark Search 16](#_Toc482882861)

[3.4.3 Focused Crawler 17](#_Toc482882862)

[3.4.4 Context Focused Crawler 17](#_Toc482882863)

[3.4.5 Info Spiders 17](#_Toc482882864)

[3.5 DESIGN IMPLEMENTATION 18](#_Toc482882865)

[Chapter 4: Result (Output) 20](#_Toc482882866)

[4.1 Indexation of EBAY (http://www.ebay.in/) 20](#_Toc482882867)

[4.2 Robots.txt file has been found. 20](#_Toc482882868)

[4.3 Sitemaps has been found in HTML not in XML. 21](#_Toc482882869)

[4.4 Good amount of TEXT found on Web Page. 22](#_Toc482882870)

[4.5 Links found while crawling. 22](#_Toc482882871)

[4.6 Main Title (specific Class) which were Crawled. 23](#_Toc482882872)

[4.7 What a Crawler See VS What a User See? 23](#_Toc482882873)

[4.7.1 What a Crawler Crawls? 23](#_Toc482882874)

[4.7.2 What a User is Able to See? 24](#_Toc482882875)

[4.8 Accessibility Overview for Web Page (using Tool). 24](#_Toc482882876)

[Chapter 5: Conclusion 25](#_Toc482882877)

[5.1 Overview of Search Engine Optimization. 25](#_Toc482882878)

[5.2 Reasons to Choose SEO as Your Career: 25](#_Toc482882879)

[5.2.1 High demand for SEO services 25](#_Toc482882880)

[5.2.2 A LOT of people have made a successful SEO career 25](#_Toc482882881)

[5.2.3 Search Engine Optimizers make Good Money 25](#_Toc482882882)

[5.2.4 Only Web–Designing MAY NOT be enough. 26](#_Toc482882883)

[5.2.5 Logical step ahead if you come from marketing or advertising 26](#_Toc482882884)

[5.2.6 Lots of Learning 26](#_Toc482882885)

[5.2.7 SEO is already recognized as a career 26](#_Toc482882886)

[5.3 Ugly face of SEO 26](#_Toc482882887)

[5.3.1 Dependent on search engines 26](#_Toc482882888)

[5.3.2 No fixed rules 27](#_Toc482882889)

[5.3.3 Rapid changes in rankings 27](#_Toc482882890)

[5.3.4 SEO requires Patience 27](#_Toc482882891)

[5.3.5 Black hat SEO 27](#_Toc482882892)

[Why SEO industry is growing? 27](#_Toc482882893)

[5.4 SEO Salary & Benefits 27](#_Toc482882894)

[5.5 Scope of SEO 28](#_Toc482882895)

[5.6 SEO Job Categories 28](#_Toc482882896)

[References: 29](#_Toc482882897)

[Annexure 30](#_Toc482882898)