Search Engine Optimization: A Survey of Current Best Practices

Niko Solihin

Follow this and additional works at: http://scholarworks.gvsu.edu/cistechlib

Recommended Citation
http://scholarworks.gvsu.edu/cistechlib/151

This Project is brought to you for free and open access by the School of Computing and Information Systems at ScholarWorks@GVSU. It has been accepted for inclusion in Technical Library by an authorized administrator of ScholarWorks@GVSU. For more information, please contact scholarworks@gvsu.edu.
Search Engine Optimization:
A Survey of Current Best Practices

By
Niko Solihin

A project submitted in partial fulfillment of the requirements for the degree of
Master of Science in
Computer Information Systems

at
Grand Valley State University
April, 2013
Search Engine Optimization: A Survey of Current Best Practices

Niko Solihin
Grand Valley State University
Grand Rapids, MI, USA
solihinn@mail.gvsu.edu

ABSTRACT
With the rapid growth of information on the web, search engines have become the starting point of most web-related tasks. In order to reach more viewers, a website must improve its organic ranking in search engines. This paper introduces the concept of search engine optimization (SEO) and provides an architectural overview of the predominant search engine, Google. This paper presents a conceptual framework for exploring various optimization guidelines, which can be categorized under four parts: keyword research, indexing, on-page optimization, and off-page optimization. Some worst practices, or “black hat” approaches, to SEO are briefly discussed. The paper concludes with the future semantic web and how it can affect optimization techniques, such as link building.

1. INTRODUCTION
This section begins with background information on Internet searches, discusses the basic tasks and primary goal of search engines, and briefly examines a research study that helps explain what this paper tries to accomplish.

1.1 Background
The web is big. In 1998, the first Google index had 26 million pages and reached the 1 billion mark by 2000. As of July 2008, that number multiplied by a thousand, or, to be exact, 1 trillion [1]. In late 2012, Google claimed to have indexed over 30 trillion unique individual live URLs on the web [2]. Search has penetrated the very fabric of our globalized society. According to comScore, more than 158 billion searches were performed worldwide each month in 2011. This number equals approximately 5.2 billion searches performed every day, and 61,000 searches performed every single second every day [3]. The way we work, play, shop, and interact have changed, and the high demand for search will continue to escalate. Profit and non-profit organizations, as well as individuals looking to have a successful presence on the web need to adapt the way they create, publish, and distribute information. Search engines are closely tied to success in the new web economy.

1.2 Search Engines
A search engine is simply a database of web pages, a method for discovering and storing information about new web pages, and a way to search that database. Therefore, in the simplest form, search engines perform three basic tasks:

1. Traverse the web and determine important words (crawling)
2. Build and maintain an index of sites’ keywords and links (indexing)
3. Present search results based on reputation and relevance to users’ keyword combinations (searching)

The primary goal is to effectively present high-quality, precise search results while efficiently handling a potentially huge volume of user queries.

1.3 Search Engine Optimization
Search Engine Optimization is the process of increasing the number of visitors to a website by achieving high rank in the search results returned by a search engine. The higher a website ranks in the results pages, the greater the chance of users visiting the website. To do this, a website uses a set of optimization methods that manipulate dozens or even hundreds of its markup elements.

A search engine results page (SERP) is the listing of results returned by a search engine in response to a keyword query, and it contains two different sections: organic and PPC (Pay-per-click). The Organic section of a SERP contains results that are not sponsored or paid for in any way, they rely strictly on search algorithms. The PPC section, on the other hand, contains text ads purchased from either Google AdWords or Microsoft AdCenter, using a bidding system to determine placements among other competing text ads. Figure 1 shows an SERP from Google.

This paper focuses on optimizing the ranking outcome of the organic result in search engines through the use of important optimization techniques. These techniques will be explored based on the conceptual framework introduced in Section 3.

1.4 Motivation for Higher Ranking
A study using heat-map testing was published in 2006 by re-search firms Enquiro, Didit, and Eyetools. Using eye-tracking devices, the study analyzed what users see and focus on when they are engaged in search activity [4]. At the time of the study, paid search listings were located in a column to the right of organic SERPs. The study showed that little attention was given to results that appear lower on the search result page (SERP). Users’ eyes also tended to be drawn to bold keywords, titles, and descriptions in the organic SERPs, while the paid search listings to the right received little attention. "The vast majority of eye tracking activity during a search happens in a triangle at the top of the search results page indicating that the areas of maximum interest create a "golden triangle" the study argues,
Figure 1: Sample Search Engine Result Page in Google, 2013.

Table 1: SERP Results Visibility

<table>
<thead>
<tr>
<th>Rank</th>
<th>Visibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100 %</td>
</tr>
<tr>
<td>2</td>
<td>100 %</td>
</tr>
<tr>
<td>3</td>
<td>100 %</td>
</tr>
<tr>
<td>4</td>
<td>85 %</td>
</tr>
<tr>
<td>5</td>
<td>60 %</td>
</tr>
<tr>
<td>6</td>
<td>50 %</td>
</tr>
<tr>
<td>7</td>
<td>50 %</td>
</tr>
<tr>
<td>8</td>
<td>30 %</td>
</tr>
<tr>
<td>9</td>
<td>30 %</td>
</tr>
<tr>
<td>10</td>
<td>20 %</td>
</tr>
</tbody>
</table>

as shown in Figure 2. Furthermore, the study provided the data shown in Table 1, displaying the varying levels of results' visibility when participants look only at the first page of a Google organic search listing.

Based on the Enquiro et al. study, we can conclude that traffic to businesses’ or individuals’ web pages is directly related to their position in the SERP. Obtaining a high rank in the search result is essential. The purpose of this paper is to deconstruct and demystify both search engines and search engine optimization techniques, to enable users to increase their search rank.

This paper introduces a helpful conceptual framework derived from surveyed literature for readers who are looking into optimizing their rankings in search engine result pages.

2. ANATOMY OF A SEARCH ENGINE

This section provides a brief architectural overview of the pre-dominant search engine, Google, and explores its various components that aid in accomplishing its major tasks: crawling, indexing, and searching. The high-level system overview is based on Sergey Brin and Lawrence Page’s original paper [5], upon which the prototype of Google was conceived. This section also discusses PageRank, a Google algorithm for measuring the importance of a web document based on its inbound link profile.

2.1 High-Level Architecture

In order to deliver fast and relevant search results, Google must first discover the hundreds of millions of web pages that exist on the World Wide Web and organize them in an efficient manner. Google utilizes special application programs called web crawlers or spiders to automatically download a copy, or cache, of a web page and follow any links it may have. After downloading, a search engine will analyze the content of each page to discover lists of words and their occurrence, structural analysis, hyperlinks, and HTML validation.

The process of crawling starts with the URL server (Figure 3) sending a list of URLs to be fetched by Googlebot, Google’s web crawler, which consists of distributed networks of high-performance computers. Fetched pages are then handed to the store server, which compresses and assigns an ID number called "docID" to each web page before storing them in a repository. Subsequently, the indexer reads the repository, decompresses the documents, and parses them. It then creates a word occurrences list called a "hit" for every document, recording the word, position in the document, capitalization, and an approximation of font size. These hit lists are then distributed into a set of "barrels", creating a partially-sorted forward index as shown in Table 2.

Additionally, the indexer parses out all the hyperlinks in every web page and stores important information about them in an anchors file. For each hyperlink parsed out, this file stores the previously assigned docID, the text of the link, as well as where each link points from and to. The URL Resolver reads the anchors file and converts relative URLs into absolute URLs and, ultimately, docIDs. It puts the anchor
text into the forward index, associated with the docID to which the anchor points. It also generates a database of links which are pairs of docIDs. The links database will be used to calculate a document’s PageRank by the Searcher, a web server responsible for handling the Google search functionality when invoked by the user.

The partially sorted forward index is then converted to an inverted index by the Sorter. This is done by taking the barrels, which are sorted by docID, and resorting them by wordID. Next, a Lexicon program takes this list and builds a new lexicon to be used by the Searcher, together with the PageRanks, to answer user queries.

Suppose we have an array of documents which has a sentence each:

```
Documents[0] = "it is what it is"
Documents[1] = "what is it"
Documents[2] = "it is a banana"
```

The inverted index of the array above is listed below in Table 3, with document numbers identifying the location of each word. A search query of "what is it" consists of the term "what", "is", and "it" and will invoke the intersection set operator, giving the result set in Figure 4.

Since these query words appear in document 0 and 1, the ID of these documents will be returned, and in Google’s case, the docID will be returned and rendered appropriately on the SERP page, sorted by rankings that PageRank calculates.

\[
\{0,1\} \cap \{0,1,2\} \cap \{0,1,2\} = \{0,1\}
\]

**Figure 4: Result Set.**

### 2.2 PageRank Algorithm

PageRank is an algorithm originally implemented by Google to measure the importance of a web document based on its inbound link profile. PageRank helps Google produce high-quality precision results. It is a purely numerical measurement and does not take into account the relevance of the linking page, or how trusted or authoritative that page might be. In a nutshell, each inbound or backlink to a web page is counted as a vote for that page, and the more backlinks a web node receives from another web node, the higher the web page will be positioned within the organic search results. Additionally, PageRank only deals with organic SERPs. The following section provides a brief overview of the original PageRank algorithm, while section 4 discusses the factors that influence link value.

#### 2.2.1 PageRank Calculation

Brin and Page define the original PageRank as follows [5]:

\[
PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right)
\]

We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. C(A) is defined as the number of links going out of page A. The PageRank of a page A is given as follows:

\[
PR(A) = (1-d) + d \left( \frac{PR(T1)}{C(T1)} + \ldots + \frac{PR(Tn)}{C(Tn)} \right)
\]

Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages’ PageRanks will be one.

To help readers understand the underlying logic of PageRank, which is still in use today, Figure 6 provides a simple diagram of the concept.

To begin with, pages are given a small amount of natural PageRank which can then be increased by increasing the

### Table 2: Forward Index

<table>
<thead>
<tr>
<th>Documents</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>the, cow, says, moo</td>
</tr>
<tr>
<td>Document 2</td>
<td>the, cat, and, the, hat</td>
</tr>
<tr>
<td>Document 3</td>
<td>the, dish, ran, away, with, the, spoon</td>
</tr>
</tbody>
</table>

### Table 3: Inverted Index

<table>
<thead>
<tr>
<th>Words</th>
<th>Doc #</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>{ 2 }</td>
</tr>
<tr>
<td>banana</td>
<td>{ 2 }</td>
</tr>
<tr>
<td>is</td>
<td>{ 0, 1, 2 }</td>
</tr>
<tr>
<td>it</td>
<td>{ 0, 1, 2 }</td>
</tr>
<tr>
<td>what</td>
<td>{ 0, 1 }</td>
</tr>
</tbody>
</table>

### Figure 5: Simple PageRank.
### 3. CONCEPTUAL FRAMEWORK

After studying primary literature on Search Engine Optimization and observing the characteristics of optimization techniques, we established a conceptual framework that offers helpful guidelines for readers as they approach this comprehensive discipline. The framework, shown in Figure 6, classifies optimization techniques under two broad categories: white hat and black hat SEO, or, techniques that search engines recommend as best practices and those which are considered unethical and manipulative. White hat optimization techniques can be further categorized by their characteristics: keyword research, indexing, on-page, and off-page optimization. Whereas black hat optimization techniques fall under two categories: content spam and link spam.

![Conceptual Framework](image)

Figure 6: Conceptual Framework.

number of backlinks pointing to the page. The amount of PageRank that a page can pass to another through its links, called passable PageRank, is predicted to be around 85% to 95% of its own PageRanks [6].

In the original formula, for pages linking to more than one other page, the passable PageRank gets divided equally among all the pages receiving those links. However, this does not hold true today, since Google has refined the original algorithm and the details of this is not widely known. In addition to outbound links, cross-linking, or the reciprocal link between two pages, requires more complex calculations. In the scenario of Figure 5, the PageRank of each page must be determined by recursive analysis, since the PageRank of Page A is dependent on f(y), the passable PageRank of Page B, which is dependent on f(x), the passable PageRank of Page A. To complicate matters more, The PageRank that Page A passes to Page C is affected by the link from Page B to Page A.

Although Google’s ability to compute reputation based on links has advanced significantly over the years, familiarity with its initial reference architecture and PageRank algorithm will help conceptualize the complexity of search engines. Furthermore, the information presented in this section helps establish the foundation for the important techniques discussed in the next section of this paper.

### 4. RECOMMENDED OPTIMIZATION METHODS

The section explores techniques that are recommended as best practices for search engines. They may be grouped as follows: keyword research, indexing, on-page, and off-page optimization. The characteristics of each optimization category will be reviewed in the following subsections.

As previously mentioned, the three important tasks of search engines are: traversing the web and determining important words (crawling), building and maintaining an index of sites’ keywords and links (indexing), and presenting search results based on reputation and relevance to users’ search terms (searching). Nearly all search engines’ operations involve keywords and, thus, keyword research forms the foundation of our conceptual framework. The result of keyword research will affect a website’s information architecture. The subsequent optimization techniques of indexing, on-page as well as off-page optimization will take keyword research further into the alteration of page elements. Above all, it is important to remember the objective of building for end users, not for search engines, in applying the methods listed in this section.

#### 4.1 Keyword Research

Keyword research is one of the most important and high-return activities in SEO, because its result will directly correlate to a site’s information architecture. Every website, whether it belongs to a business, organization, or an individual, has a customer base associated with it. Within each customer base, there is a keyword demand, which, if examined properly, will enable website owners to predict shifts in demand, respond to changing market conditions, and ensure that they are producing the products, services, and content that customers are actively seeking. Since every word and phrase that is typed into search engines is recorded and often available to keyword research tools, website owners will be able to gain a general comparison of high-demand keywords within their niche customer base. Moreover, they will be able to understand the value of a keyword by doing simple research, making some hypotheses, testing, and repeating the cycle while continuously monitoring the site’s click-through (CTR) and conversion rate.

Although there are many tools and methods that are usually used by SEO practitioners, Google’s AdWords and Trends are common starting points for keyword research and they will be explored here. AdWords is aimed at helping Google’s paid search customers, but it is often used to obtain information for organic search. A brief overview of AdWords and Trends will be provided. However, specifics about its usage is beyond the scope of this paper.

##### 4.1.1 Google AdWords

The Google AdWords Tool ([https://adwords.google.com/o/KeywordTool](https://adwords.google.com/o/KeywordTool)) consists of two tools: Keyword Tool and Traffic Estimator. Keyword Tool provides related terms, search volume estimates, search trends, and ad cost estimates for any keyword or URL that a user enters, while Traffic Estimator enables website owners to get estimates of traffic on different keywords (see Figure 7). Essentially, SEO practitioners must pay careful attention to the numbers circled in the diagram below - for example, how competitive a keyword is, what its global/local monthly searches are, and what its potential click-through rate is.
These numbers indicate whether a keyword is too broad or too narrow and whether it will bring any advantage to a website.

4.1.2 Google Trends

Google Trends (www.google.com/trends/) enables SEO practitioner to compare two or more search terms to see their relative popularity, as well as seasonality/trending over a specific time range. Additionally, it will also provide location information of a keyword’s market. Figure 8 compares the trend of keywords "pepperoni pizza" to "sausage pizza" in the United States within the past year. Google Trends shows that pepperoni pizza has more interest than sausage pizza, especially in Oklahoma, Ohio, and Kansas.

4.2 Indexing

This section examines methods related to the accessibility of a website to search engines spiders. Ensuring that a website’s content and links can be found and crawled by search engines is an important step toward creating visibility in search results. First, the common reasons why pages may not be indexable are given. Next, methods that can be used to increase the chance of website crawlability will be discussed. Last, the use of the site: search operator will be explained to examine which pages of a website have been crawled by Googlebot. Using the techniques in this section, website owners will be able to determine whether information architecture can be improved for more crawlability.

4.2.1 Crawlable Content and Link Structure

To rank well in SERP, content material for end users must be in HTML text form. Despite the significant progress in technology, crawlers are by nature optimized to handle plain HTML text instead of images and third party multimedia platforms such as Flash. From a relevance perspective, it is difficult for search engines to analyze text content that are embedded inside of an image or Flash file. Crawlers will only be able to read the text in a filename for Flash files and text in the filename, title, and alt attribute for images...Therefore, indexing preference is almost always given to HTML-based text files.

In terms of links, crawlers give preference to links that use descriptive keywords that described linked content instead of, for instance, using a simple "click here" to label anchor links on pages. Furthermore, common reasons why pages may not be reachable to spiders such as Googlebot are listed below:

- Links in submission-required forms
- Links in Javascript AJAX calls
- Links in Flash, Java, or other plug-ins
- Links in PowerPoint and PDF files
- Links pointing to pages blocked by the meta robots tag rel="NoFollow", or robots.txt
- Links on pages with many hundreds or thousands of other links
- Links in frames or iFrames
- Some of the points above will be discussed in other sections of the paper.

4.2.2 Index Optimization

Besides the obvious method of linking a newly created page from an already indexed site, there are other methods discussed in this section that will ensure new pages are crawled and indexed. They consist of submitting new content directly to Google and using sitemap.xml for the main site and images.

According to Matt Cutts, the head of the webspam team, Google is able to discover and index new content within 24 hours [7]. However, there may be instances where a manual submission of sites is necessary. Page owners can submit an unindexed page directly to Google at www.google.com/submityourcontent/ Another way to complement search engines’ normal, link-based crawl is the use of a protocol known as XML Sitemaps. Using the Sitemaps protocol, page owners can supply the search engines with a list of all the pages they would like them to crawl and index. Additionally, sitemaps can also increase visibility of images that are listed within. Sitemaps use a simple XML format defined in detail at http://www.sitemaps.org and must reside within the highest-level directory, or the root directory.

4.2.3 Inspecting Indexed Pages

Google supports a number of advanced search operators that can be used to help diagnose SEO issues. The site: operator in particular can be used to execute a domain-restricted search (.com, .net, or .org) and narrow a search to one or more specific subdomain or directories. To show approximately how many URLs are indexed by Google, one may use site:example.com. Additionally, to show indexed URLs from a specific directory, subdomain, or top-level domain (TLD), one may use site:example.com/directory, site:subdomain.example.com, or site:org respectively.

4.3 On-page Optimization

On-page optimization refers to the alteration of numerous on-page elements to help search engine spiders determine what the page is about and how it may be useful for users. During this optimization stage, it is important to remember the notion of building primarily for end users, not for search engines. Every element on the page - document text, image, or links - must ultimately provide value for end users.

In order to drive traffic, the targeted keywords from the previous stage must be included in the pages to be optimized. Keyword relevance must be aligned with the usability of a page from a human perspective, thus the practice of excessively placing keywords into every HTML tag possible is strongly discouraged and can, in fact, lead to penalization from search engines. This section explores some of the more prominent places where site owners can strategically place targeted keywords. It provides a brief suggestion related to keyword placement in specific tags, according to the literature surveyed.

4.3.1 Title Tags

For keyword placement, title tags are the most critical element for search engine relevance. Located in the <head> section of an HTML document, it is the only piece of "meta" information about a page that influences relevancy and ranking. A page’s title tag content must align with its actual visible content. Aside from incorporating keyword phrases in the title tag, a brief list of best practices that yield the most search engine benefits is provided below:
Figure 7: Google AdWords.

Figure 8: Google Trends.
4.3.2 Meta Description Tags

The `<meta name="description" content="...">` is primarily used to describe the content of the page accurately and succinctly, and to serve as a short advertisement text in the search result (see Figure 9).

Aside from incorporating keyword phrases in the description, a brief list of best practices that yield the most search engine benefits is provided below:
- Describe brand and content honestly
- Limit length to 165 characters (including spaces)
- When needed, use | (pipe), -, :, or > as dividers
- Target searcher intent by using verbs in keywords

4.3.3 Heading Tags

Heading tags are designed to indicate a headline hierarchy in a document. An `<h1>` tag might be considered the headline of the page as a whole, whereas `<h2>` tags would serve as subheading, `<h3>` tags as tertiary-level headlines, and so forth. Search engines have shown preferences for keywords appearing in heading tags, especially the `<h1>` tag. SEO professionals have claimed that it is safe to use the content of `<title>` tag as heading tags.

4.3.4 Other On-Page Optimization Techniques

This subsection discusses other important guidelines regarding on-page optimization.

Beside the aforementioned important keyword placements, other common places for keywords that will yield most search engine benefits are links and texts near the top of the page, URLs, the body, alt and title attributes of `<img>`, and image filename.

As far as URL is concerned, search engines have shown preference for short, readable URLs with targeted keywords included. The use of hyphen as a separator in URLs is encouraged. Complex query parameters can result in lower overall ranking and indexing.

Another important technique is that of redirection. There are two major types of redirects that are typically used to indicate when content has moved from one location to another: 301 Moved Permanently and 302 Moved Temporarily. Crawlers treat these differently. A 302 HTTP status code instructs the crawler to not pass any PageRank from the old to the new page since the move is temporary. On the other hand, a 301 redirect will ensure that PageRank and any other page metrics will be transferred to the new page.

Pagination can also be optimized using the `<link rel="next" href="url-of-next-page">` and `<link rel="prev" href="url-of-previous-page">` markup in the `<head>` section of every paginated page. This allows crawlers to index paginated pages more thoroughly, especially if a section of a website is heavily paginated. For instance, if a section of a website is divided into 3 pages, the first page will only contain a rel="next" since there is not any page before it, the second page will contain both rel="prev" and rel="next", and the last page will only contain a rel="prev" since there is not any page after it.

4.4 Off-page Optimization

Off-page optimization refers to the link building process that influences how search engines rank a web page. It is clear that links play a critical role in ranking — it is the decisive vote in SERP. This section is divided into two parts: link building and social signals.

4.4.1 Link Building

Link building is the practice of actively marketing a website with the intent to obtain links from other sites. Link building is a fundamental part of SEO and is an ongoing aspect of marketing a website. Search engines use links not only to discover web pages, but also to determine rankings and how frequently and deeply a website is crawled. Google measures the value of links by PageRank and a few other factors:

- **Anchor Text**
  The clickable part of a link is a strong ranking element that helps search engines understand what the page receiving the link is about.

- **Relevance**
  Fundamentally, relevance refers to more weight that is given to links that originate from sites or pages on the same topic, or sites or pages that are closely related to the owner’s site.

- **Authority**
  Although it is not clear how Google carries out its topical relevance analysis, authority is closely related to the notion of link analysis through the hubs and authorities model. Hubs are sites that link to most of the important sites relevant to a particular topic while authorities are sites that are linked to by most of the sites relevant to a particular topic [8].

- **Trust**
  It is likely that Google utilizes some form of TrustRank, a purely algorithmic approach to determine the trust of a site [9]. The trust level of a site is based on how many clicks away it is from one or more seed sites — a set of highly trusted sites selected by manual human review. A site that is one click away accrues a lot of trust; two clicks away, a bit less; three clicks away, even less; and so forth.

There are many different link building strategies that are beyond the scope of this paper. However, these strategies usually share some common characteristics:
• Use some form of link baiting
  Building content that plays to the emotions of potential linkers - for example, a website with useful or entertaining content.

• Provide unique and quality information to users
  Creating quality reference material that authoritative site owners find to be of value to their audience.

• Leverage business relationships
  For instance, having a network of distributors and resellers for a company’s product that link back to the company’s page as a standard term in the business agreement.

4.4.2 Social Signals

In addition to inbound links, social media and user engagement have become significant factors in ranking, as search engines learn to utilize these metrics more effectively. Social media sites such as Facebook, Google+, and Twitter are venues that businesses, organization, or individual can take advantage of to reach out to their audience.

As of 2011, it was confirmed that both Google and Bing treat shared links on walls and Facebook Fan pages as votes for the website being shared. More specifically, Google clarified that it treats Facebook links in the same way that it treats tweeted links [10]. This suggests that Facebook and Twitter can play an important role in a link-building campaign. They are an excellent way to develop exposure and share links of value by using some of the link building approach in the previous section.

In short, excellent link building comes from a simple idea: create great product or content, tell everyone about it, and motivate them to share. Additionally, as part of their ongoing off-page optimization strategy, site owners must always consider participating in social media communities, providing an engaging user experience, offering unique and differentiated content, and building a strong brand.

5. MANIPULATIVE OPTIMIZATION METHODS

While it is true that a large part of Search Engine Optimization involves manipulating the site’s content and HTML tags, there is a limit of how far a page can be manipulated before it raises a red flag for search engines. Certain optimization techniques have been categorized as "Black Hat" or spamdexing techniques, that is, techniques that are generally used in an unethical manner to get higher search rankings.

Neglecting the rules set by the search engine will lead to penalties, such as a lower search ranking or, in some cases, a permanent ban from the search engine. In 2006, Google banned the BMW Germany website for using a "doorway" page — a page designed to present different content to web crawlers than to human audiences. Google states that it may "...temporarily or permanently ban any site or site authors that engage in tactics designed to distort their rankings or mislead users in order to preserve the accuracy and quality of our search results." [11]

It is important for SEO practitioners to be aware of techniques that are considered unethical and the impact they have on search engine ranking. This section examines spamdexing based on two types of spam: content spam and link spam.

Some spamdexing tactics that are still commonly practiced for each type of spam will be explored.

5.1 Content Spam

This category of spamdexing involves techniques that modify the logical view of a page to search engines, such as keyword stuffing, doorway pages, and invisible elements.

5.1.1 Keyword Stuffing

Keyword stuffing involves placing excessive amounts of keywords into the page content and <meta name="keywords" content="..."> tag, in such a way that they detract from the readability and usability of a given page. To boost the page’s ranking in the SERP, both relevant and irrelevant keywords are inserted and occasionally repeated. Frequently searched or high-commercial-value spam keywords, such as mortgage, poker, texas hold ’em, porn, are inserted and repeated in the page <meta> tag or content.

Common methods of keyword stuffing in page content include hiding keywords on the page by making the text the same color as the background, hiding keywords in comment tags, exploiting CSS z-index or position: absolute properties to get keywords behind or off-page, and overfilling alt tags with long strings of keywords.

Although it was never mentioned explicitly, search engine experts believe that Google no longer uses keywords in the <meta> tag as a ranking factor due to past abuses of this tag. Bing, on the other hand, still takes it into account as a signal for "spammy" content. Furthermore, experts believe that proper usage of keywords in the <meta> tag has no ranking benefits, whereas abusing it will lead to penalized rankings with both Google and Bing [12]. Keyword stuffing is an unethical way to optimize ranking that deceives search engines and its use is strongly discouraged.

5.1.2 Doorway Pages

Also known as a "bridge page" or "gateway page", a doorway page is a web page filled with keyword-rich content that does not deliver any useful information other than a redirect to an external site. The primary goal of doorway pages is to gain high rankings for multiple keywords and phrases — to capture search engine traffic through those keywords and send that traffic to a highly commercial web page (see Figure 10).

Typically, doorway pages are created using optimization tools. These tools automatically create separate pages, sometimes up to hundreds of pages, for each keyword or phrase and set up an instant redirect to the intended page using <meta http-equiv="refresh" content="0;URL=http://example.com/">. Search engines have combated this by not indexing pages that contain instant meta redirects. However, black hat optimizers have fought back with a variety of other techniques, including the use of Javascript, PHP, and other server-side technologies.

5.1.3 Cloaking

Slightly similar to the doorway pages method is cloaking. Cloaking involves using user agent or IP address detection to recognize incoming visitors as either search engine spiders or users, and then delivering different content to the spiders than that seen by human users. It is used to gain high rankings on multiple major search engines. The pages intended for users will contain images, styling, and other de-
sign elements, whereas the pages for search engines spiders are typically text only. IP address and User-Agent HTTP Header detection are key components to cloaking, since the IP addresses and User-Agent for most of the major search engine spiders are well known. Figure 11 illustrates cloaking.

5.2 Link Spam

This category of spamdexing involves generating links that do not provide end-user value. These links exist for the sole purpose of inflating search engine rankings. Two forms of link spam will be discussed in this section: link farms and comment spam.

5.2.1 Link Farms

A link farm is a group of highly interlinked websites formed for the purpose of inflating link popularity or PageRank. Web pages utilizing manipulative search engine tactics often participate in link farms or reciprocal link exchanges. Instead of acknowledging relevant content from other authors in the same field by mutual linking, link farms participants inserts cross links that point to irrelevant information that provide no value for the end-user. Link farms are typically created automatically by tools that can generate hundreds of links with little effort. Seeing this as an advantage for lucrative income, many spamdexing businesses boost web page rankings instantly by employing this tactic. Search engines such as Google and Bing have identified specific attributes associated with link farm pages and will entirely remove domains that they suspect are collaborating in a link farm or reciprocal link exchange scheme. Figure 12 shows the concept of link farm with arrows indicating a cross link pair [13].

5.2.2 Comment Spam

Comment spams are comments posted for the purpose of generating an inbound link to another site that will reap the benefit of having higher PageRank. They are typically
posted on wikis, blogs, and guestbooks by tools that are able to comment on hundreds of blogs at a time. The most commonly used method to combat comment spam is the rel="nofollow" attribute on anchor tags in the comment section. Furthermore, third-party comment management systems such as Disqus (www.disqus.com) utilize iFrames or Javascript AJAX requests to dynamically render comments, making sure they are non-crawlable by search engines. Nevertheless, there are still many blogs and wikis that are not using these solutions, making them targets for comment spambots.

6. FUTURE OF SEARCH

Larry Page, co-founder and CEO of Google, once described the "perfect search engine would understand exactly what you mean and give you back exactly what you want." The word "mean" or "semantic" is very important in the ever-changing discipline of Information Retrieval. This section explores Semantic Web, how it relates to search. An example of Semantic Search technology, the Google Knowledge Graph, is briefly mentioned.

6.1 Semantic Web

In 2001, the inventor of the World Wide Web, Tim Berners-Lee and a few other computer scientists, emphasized the significance of expressing meaning through the use of Semantic Web technologies in order to bring elaborate, precise automated searches. Berners-Lee defines Semantic Web as the "extension of the current Web in which information is given the well-defined meaning, better enabling computers and people to work in cooperation." [14] The Semantic Web technologies will "enable explicit, unambiguous, and shared definitions of domain terms and relations (for humans and machines to interpret) as well as a global system for identification and reuse." [15] Furthermore, Semantic Web will bring the Web closer to its 3.0 phase, which focuses on meanings and connecting knowledge to enable more relevant, useful, and enjoyable user experience of the Internet.

6.2 Semantic Search

Semantic search promises to produce precise answers to user’s queries by taking advantage of availability of explicit semantics of information in the semantic web. In other words, instead of relying on dictionary or index retrieval based on keywords, some degree of artificial intelligence will be utilized to understand the searcher’s intent and the contextual meaning of the query terms. Major players in the search engine world such as Google and Bing have started incorporating some aspects of semantic search and will continue to make significant advancements in this area. Semantic search uses the science of meaning in natural language to produce results that are relevant to that specific user. The goal is to deliver the information queried by a user rather than having a user sort through a list of loosely related keyword results.

6.3 Google Knowledge Graph

Incorporated into its search systems in 2012, the Google Knowledge Graph is a knowledge base used by Google to enhance its search engine’s search results with semantic-search information gathered from a wide variety of sources [16]. From the searcher’s point of view, the Knowledge Graph enables Google to become an intelligent information assembler. According to Amit Singhal, the head engineer of Google’s core ranking team, the Knowledge Graph derived its information not only from public sources such as the CIA World Factbook, Freebase and Wikipedia, but also many other large scale information databases. It currently contains more than 500 million objects, as well as more than 3.5 billion facts about and relationships between these different objects [17, 18].

As an information assembler, the searcher does not need to navigate to websites on listed on SERPs to gather useful, relevant information. Anything that is closely related to what the searcher intended will be displayed, leading to new, unexpected discoveries. For instance, the query term "famous jazz composers" will present a carousel UI unit with photos and names of the different composers. Clicking each name will present a wealth of useful, relevant information, such as key facts, albums recorded, and popular songs composed (see Figure 13).

Additionally, the Knowledge Graph also considers the searcher’s personal information to which Google has direct data access, such as via GMail. Precise information that is tailored to that particular user will be laid out based on what Google knows about that individual. For instance, the query term "my flight" will organize flight confirmation emails for any upcoming trips in a beautifully easy-to-read manner on the search results page [19] as shown in Figure 14.

Figure 14: Knowledge Graph displaying user’s data.
Figure 13: Knowledge Graph in Google SERP.
7. REFERENCES


