

Search Engines

Chapter 20

Objectives

1

The **History** and **Anatomy**
of Search Engines

2

Web **Crawlers** and Scrapers

3

Indexing and Reverse
Indexing

4

PageRank and Result
Order

5

White-Hat Search Engine
Optimization

6

Black Hat Search
Engine Optimization

Section 1 of 6

THE HISTORY AND ANATOMY OF SEARCH ENGINES

Google

It's now a word

The impact of search engines is so pronounced that *The Oxford English Dictionary* now defines the verb **google** as

Search for information about (someone or something) on the Internet using the search engine Google.

This shift in the way we retrieve, perceive, and absorb information is of special importance to the web developer since search engines are the medium through which most users will find our websites.

Before Google

Not that long ago

In the days before Google there was no capacity to search the entire WWW. Users would learn about websites by following a link from an email, a message board, or other site.

By 1991 sites dedicated to organized lists of websites started appearing, often created and curated by the Internet Service Providers who wanted to provide added value to their growing clientele.

These **web directories** categorized websites into a hierarchy and still exist today.

Before Google

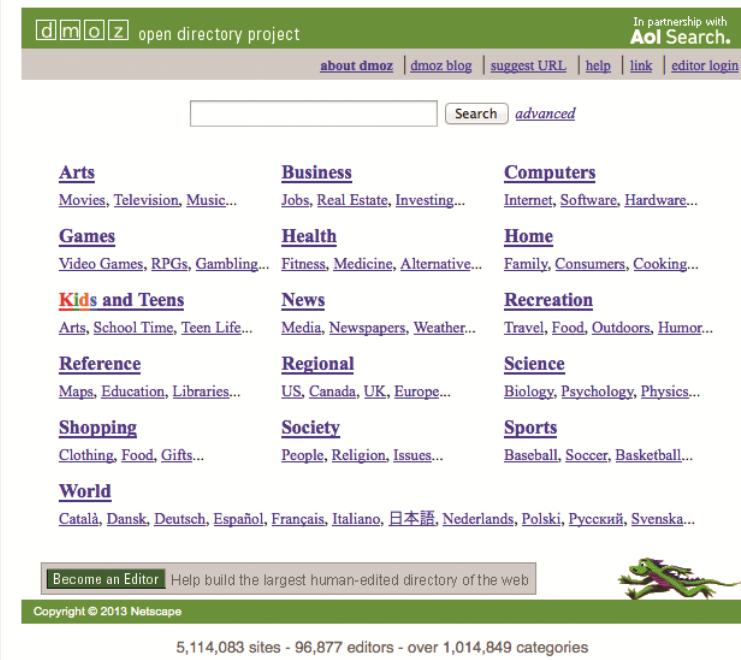
Web Directories

To be added to a web directory, one would have to submit a request, often by email.

In curated directories the webmasters would then decide whether or not to list you, and if so, where. Many sites took it upon themselves to censor which sites would be listed.

The Open Directory Project (dmoz.org) has a more open philosophy.

Yahoo was this way, not sure if it still is



Before Google

Early crawlers

In 1993 **web crawlers**, the first component of search engines, started appearing.

These crawlers could download a page and parse out all the links to other pages (backlinks).

Meanwhile, in 1996, graduate students at Stanford, Lawrence “Larry” Page, and Sergey Brin began working on a crawler. They incorporated as Google Inc. in 1998, and by June 2000 Google had grown their index to over 1 billion URLs (by 2008 it was 1 trillion). Today I have no idea....

Search Engine Overview

Lots of components

Search engines consist of several components:

- **Input agents:** web crawlers surf the WWW requesting and downloading web pages
- **Database engine:** manages the URLs and the agents in general
- **The query server:** handles requests from end users

In practice, these components are distributed although conceptually they can be thought of as services on the same machine

Section 2 of 6

WEB CRAWLERS AND SCRAPERS

Crawlers

Spiders, robots, wanderers

Web crawlers refer to a class of software that

- downloads pages
- identifies the hyperlinks, and
- adds links to a database for future crawling

A crawler can be written to be autonomous, so that it populates its own list of fresh URLs to crawl, but is normally distributed across many machines and controlled centrally.

Crawlers

Doesn't do much by itself...

```
class Crawler {
    private $URLList;
    private $nextIndex;
    function __construct(){
        $this->nextIndex=0;
        $this->URLList = array("http://SEEDWEBSITE/");
    }
    private function getNextURLToCrawl(){
        return $this->URLList[$this->nextIndex++];
    }
    private function printSummary(){
        echo count($this->URLList). " links. Index:".
            $this->nextIndex."<br>";
        foreach($this->URLList as $link){
            echo $link."<br>";
        }
    }
    // THIS CAN BE CALLED FROM LOOP OR CRON
    public function doIteration(){
        $url = $self->getNextURLToCrawl();
        if (robotsDisallow($url))
            return;
        echo "Crawling ".$url."<br>";
        scrapeHyperlinks($url);
        $self->printSummary();
    }
}
```

I think they're skipping hard part, the parsing

LISTING 20.1 Simple crawler class in PHP

Crawlers

Be Polite!

In the early days of web crawlers there was no protocol about how often to request pages, or which pages to include, so some crawlers requested entire sites at once, putting stress on the servers.

Martijn Koster, the creator of [ALIWEB](#), drafted a set of guidelines enshrined as the **Robots Exclusion Standard** that help webmasters block certain pages from being crawled and indexed

inadvertent denial of service attack

Robots Exclusion Standard

Be Polite!

The Robots Exclusion Standard is implemented with plain text files named **robots.txt** stored at the root of the domain.

Robots.txt has two syntactic elements

1. user-agent we want to make a rule for (the special character * means all agents)
2. One Disallow directive per line to identify patterns.

Regular expressions are not supported!!

Robots Exclusion Standard

Robots.txt

Could be important
for web programmers
of big sites

The Robots Exclusion Standard is not a layer of authentication or security.

Some malicious bots will not obey the directives and purposefully seek out materials specifically disallowed in **robots.txt**.

```
User-agent: googlebot  
Disallow:
```

Allow all , could just have no file

```
User-agent: funbot  
Disallow: /secret/
```

Please don't search in the secret directory

```
User-agent: *  
Disallow: /
```

Please don't search here at all.

LISTING 20.2 Robots.txt to allow googlebot full access, allow funbot partial access, and block all other bots

Prioritization

Don't download the entire site at once!

Prioritization - ranking the uncrawled URLs, using techniques like **PageRank**

A combo of PageRank and a timestamp of the last time a domain was accessed is the start of a prioritization policy.

Scrapers

“Readers” of my website

Scrapers are programs that identify certain pieces of information from the web to be stored in databases.

Sometimes combined with Crawlers. There are several classes of Scraper:

- URL Scrapers
- Email Scrapers
- Word Scrapers
- Media Scrapers

23% of all internet traffic

URL Scrapers

Links: The “threads of the web”

URL Scrapers identify URLs inside of a page by seeking out all the <a> tags and extracting the value of the href attribute. This can be done through string matching, seeking the <a> tag, or more robustly by parsing the HTML page into a DOM tree

```
$DOM = new DOMDocument();  
$DOM->loadHTML($HTMLDOCUMENT);  
  
$aTags = $DOM->getElementsByTagName("a");  
foreach($aTags as $link){  
    echo link->getAttribute('href')." - ".$link->nodeValue."<br>";  
}
```

LISTING 20.3 PHP scraper script to extract all the hyperlinks and anchor text

Email Scrapers

Not necessarily evil

Email scrapers harvest email accounts by seeking out **mailto:** in the href attribute of a link.

A slight modification of our URL Scraper is all that's needed.

```
foreach($aTags as $link){
    $mailpos=strpos($link->getAttribute('href'),"mailto:");
    if($mailpos !== false){
        echo substr($link->getAttribute('href'),$mailpos+7)."<br>";
    }
}
```

LISTING 20.4 Portion of a PHP email harvesting scraper

Word Scrapers

That which allows us to search

A **Word scraper** may want to parse out is all of the text within a web page.

Words are the most difficult content to parse, since the tags they appear in reflect how important they are to the page overall.

- Words in a large font are surely more important than small words at the bottom of a page.
- Words that appear next to one another should be somehow linked while words that are at opposite ends of a page or sentence are less related.

Also consider the importance of words, for example the title of the page

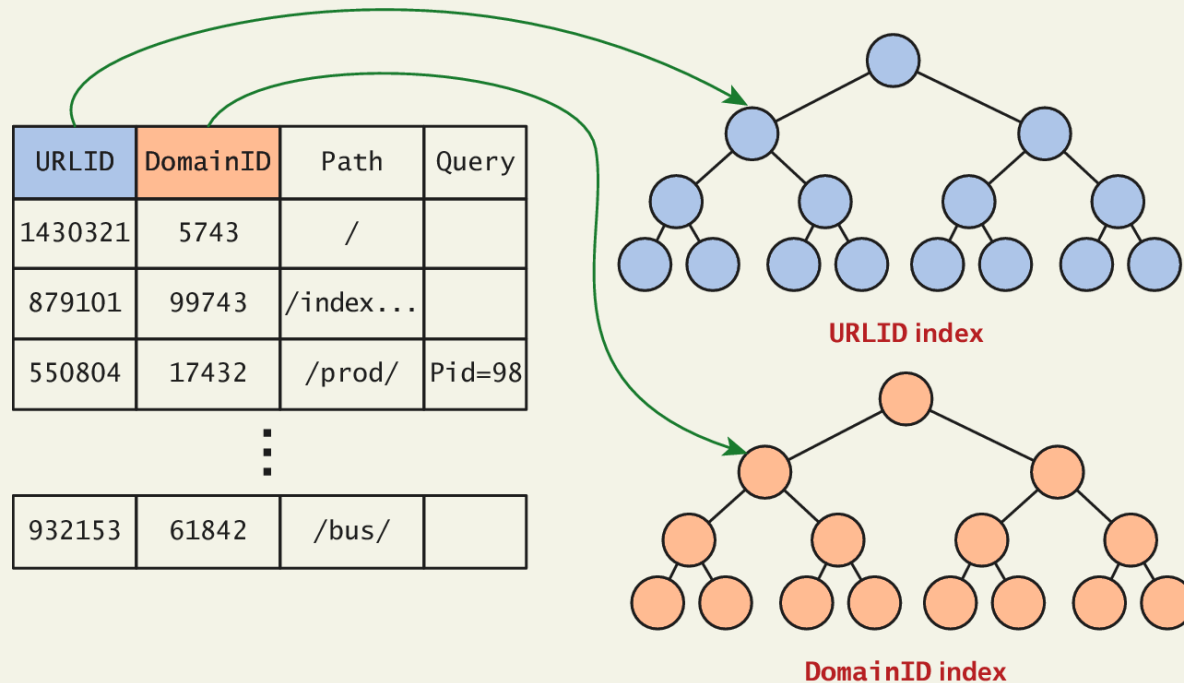
Section 3 of 6

INDEXING AND REVERSE INDEXING

Indexing and Reverse Indexing

That which makes trillion URL searches possible

To understand indexing, consider what a crawler and a scraper might identify from a web page and how they might store it.



Indexing and Reverse Indexing

That which makes trillion URL searches possible

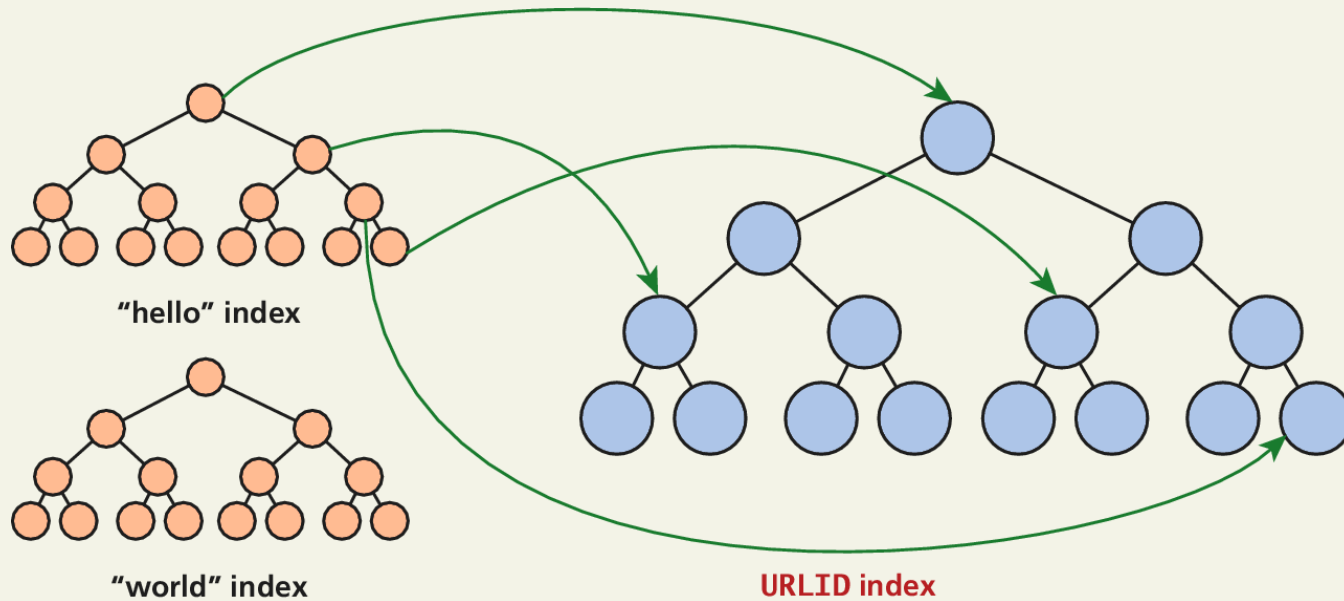
A **reverse index** essentially indexes the words, rather than the URLs.

The mechanics of how this is done are not standardized, but generally word tables are created (for every word found in pages) so that each word can be referenced by a unique integer, and indexes of these references can be built for faster searches.

Since there are tens of thousands of words, and each word might appear in millions of web pages, the demands on these indexes far exceed what a single database server can support.

Indexing and Reverse Indexing

That which makes trillion URL searches possible



Section 4 of 6

PAGERANK AND RESULT ORDER

PageRank

Bringing order to big data

PageRank is an algorithm, published by Google's founders in 1998.

According to the authors, PageRank is

a method for computing a ranking for every web page based on the graph of the web.

The *graph of the web* being referred to looks at the hyperlinks between web pages, and how that creates a *web* of pages with links.

- Sites with thousands of backlinks are surely more important than sites with only a handful of backlinks

PageRank

What, there's math?

The simplified definition of a site n 's PageRank is

$$PR(n) = \sum_{v \in B_u} \frac{PR(v)}{N_v}$$

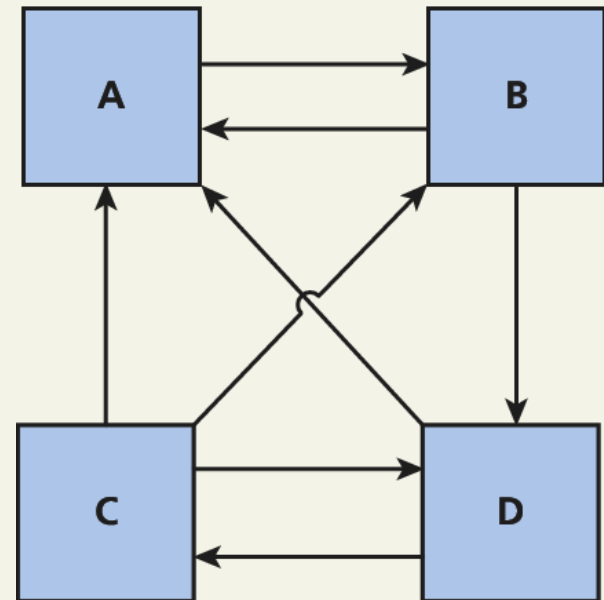
$PR(n)$, is determined by collecting every page v that links to n ($v \in B_u$), and summing their PageRanks $PR(v)$ divided by the number of **links out** (N_v).

PageRank

An example

To begin, assign the default rank to all pages:

$$PR(A) = PR(B) = PR(C) = PR(D) = \frac{1}{4}$$



PageRank

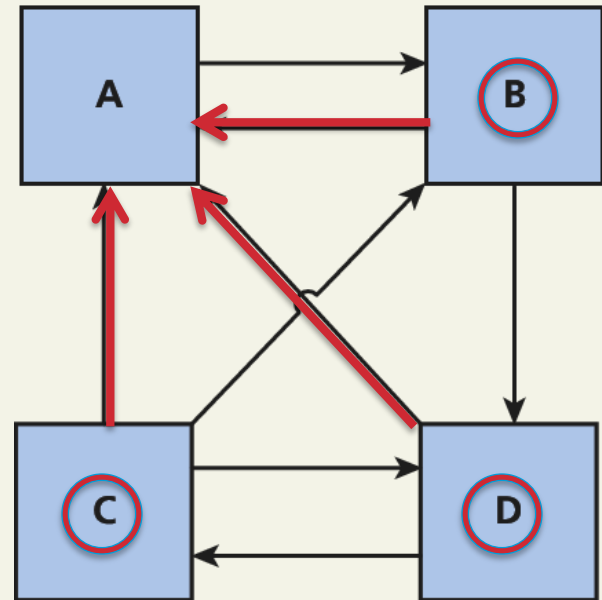
An example

Calculate the updated PageRank for A.

$$PR(A) = \sum_{v \in B_A} \frac{PR(v)}{N_v}$$

$$PR(A) = \frac{PR(B)}{N_B} + \frac{PR(C)}{N_C} + \frac{PR(D)}{N_D}$$

$$PR(A) = \frac{1/4}{2} + \frac{1/4}{3} + \frac{1/4}{2} = \frac{1}{3}$$



PageRank

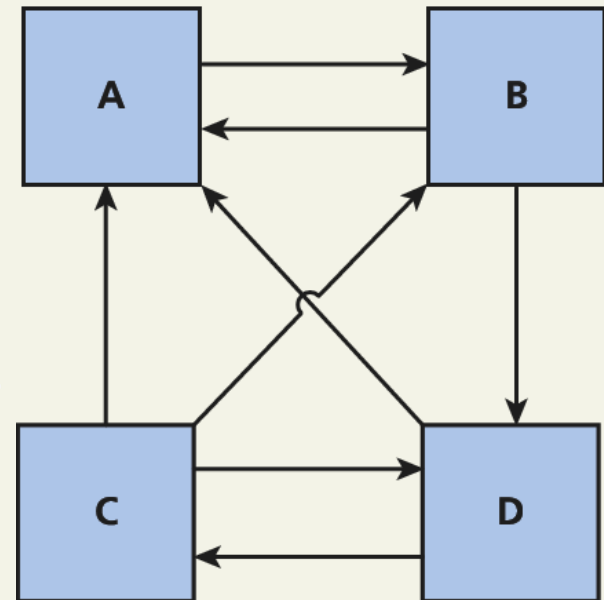
An example

Calculate the updated PageRank for the others

$$PR(B) = \frac{PR(A)}{N_A} + \frac{PR(C)}{N_C} \Rightarrow \frac{1}{4} + \frac{1/4}{3} \Rightarrow \frac{1}{3}$$

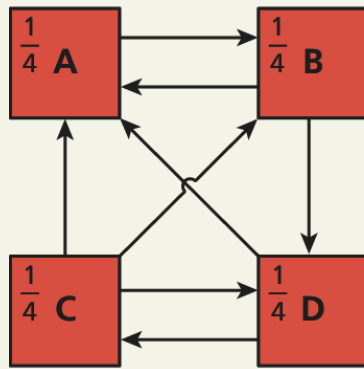
$$PR(C) = \frac{PR(C)}{N_C} \Rightarrow \frac{1/4}{2} \Rightarrow \frac{1}{8}$$

$$PR(D) = \frac{PR(B)}{N_B} + \frac{PR(C)}{N_C} \Rightarrow \frac{1/4}{2} + \frac{1/4}{3} \Rightarrow \frac{5}{24}$$

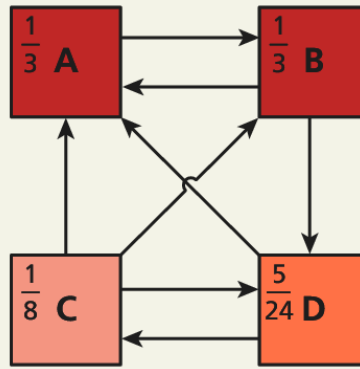


PageRank

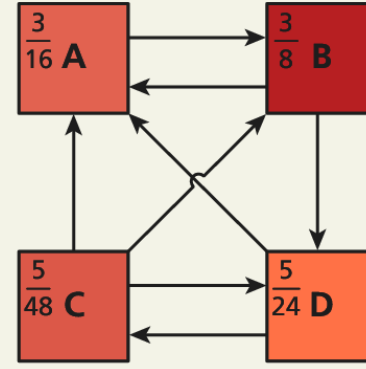
Visualized



Iteration 0



Iteration 1



Iteration 2

PageRank

Visualized

Modern ranking algorithms take much more into account than simple backlinks.

- Search History
- Geographic Location
- Authorship
- Freshness of the pages
- Other inputs...

Section 5 of 6

WHITE-HAT SEARCH ENGINE OPTIMIZATION

Search Engine Optimization

White-Hat

Search engine optimization (SEO) is the process a webmaster undertakes to make a website more appealing to search engines, and by doing so, increases its ranking in search results for terms the webmaster is interested in targeting.

- For many businesses the optimization of their website is more important than the site itself.
- Sites that appear high in a search engine's rankings are more likely to attract new potential customers, and therefore contribute to the core business of the site owner.

Search Engine Optimization

White-Hat

An entire area of research into SEO has risen up and can be broken down into two major categories:

- **White-hat SEO** that tries to honestly and ethically improve your site for search engines, and
- **Black-hat SEO** that tries to game the results in your favor.

Title Tags

White Hat Technique

The <title> tag in the <head> portion of your page is the single most important tag to optimize for search engines.


The content of the <title> tag is how your site is identified in search engine results as shown

Fundamentals of Web Development

<http://funwebdev.com>

The companion site for the upcoming textbook Fundamentals of Web Development from Pearson. Fundamental topics like HTML, CSS, JavaScript and ...

- make it unique on each page of your site
- Include enough keywords to make it relevant



But not too many as to water it down.

Meta Tags

White Hat Technique

Early search engines made significant use of meta tags, since indexing meta tags was less data-intensive than trying to index entire pages.

- Description
- Keywords
- Robots
- http-equiv

Meta Tags

White Hat Technique

The **description** meta tag contains a human-readable summary of your site.

```
<meta name="description" content="The companion site for the upcoming textbook Fundamentals of Web Development from Pearson. Fundamental topics like HTML, CSS, JavaScript and" />
```

The **keywords** meta tag allows a site to summarize its own keywords (normally ignored nowadays)

```
<meta name="keywords" content="Web Development, HTML5, CSS, JavaScript, PHP, MySQL, LAMP, Security, Search Engines, ... " />
```

Http-Equiv Meta Tags

White Hat Technique

Tags that use the http-equiv attribute can perform HTTP-like operations like redirects and set headers.

To indicate that a page should not be cached:

```
<meta http-equiv="cache-control" content="NO-CACHE">
```

To redirect to <http://funwebdev.com/destination.html> after five seconds.

```
<meta http-equiv="refresh" content="5;URL=http://funwebdev.com/destination.html">
```

Robot Meta Tags

White Hat Technique

We can control some behavior of search engines through meta tags with the name attribute set to robots. The content for such tags are a comma-separated list of INDEX, NOINDEX, FOLLOW, NOFOLLOW

To include a description and tell robots to index the site, but not to count any outbound links toward PageRank algorithms:

```
<meta name="description" content="Share your vacation photos with  
                                friends!" />  
<meta name="robots" content="INDEX, NOFOLLOW" />
```

LISTING 20.5 Meta-tag examples for a photo sharing site

URLs

White Hat Technique

Search engines must by definition download and save URLs since they identify the link to the resource.

Bad SEO URLs

work just fine for programs but cannot be read by humans.

/products/**index.php?productID=71829**

This can be improved by adding

- descriptive path components and
- descriptive file names

Ever search for
a particular
item to buy?

Good URLs

White Hat Technique

If product 71829 is an air filter, for example, then a URL that would help us identify that this is a product in a category would be

`/products/AirFilters/index.php?productID=71829`

A step further would be to add the name of the filter in the URL in place of the product's internal ID.

`/products/AirFilters/BudgetBrandX100/`

Vs the original

`/products/index.php?productID=71829`

Site Design

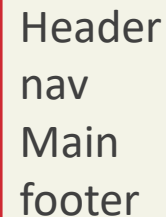
White Hat Technique

Sites that rely heavily on JavaScript or Flash for their content and navigation will suffer from poor indexing.

If your site includes a hierarchical menu, you should nest it inside of `<nav>` tags to demonstrate **semantically** that these links exist to navigate your site.

Links in a website can be categorized as:

- navigation,
- recurring, and
- ad hoc.



Header
nav
Main
footer

SiteMaps

White Hat Technique

A formal framework that captures website structure is known as a **sitemap**. Using XML, sitemaps define a URL set for the root item, then as many URL items as desired for the site.

```
<?xml version="1.0" encoding="utf-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://funwebdev.com/</loc>
    <lastmod>2013-09-29</lastmod>
    <changefreq>weekly</changefreq>
    <priority>1.0</priority>
  </url>
</urlset>
```

LISTING 20.7 Single page sitemap

Anchor Text

White Hat Technique

One of the things that is definitely indexed along with backlinks is the **anchor text** of the link.

- In the early web, many links said *click here*
- These days, that use of the anchor text is not encouraged, since it says little about what will be at that URL

Links to a page of services should read “*Services and Rates*,” since that anchor text has keywords associated with the page.

Images

White Hat Technique

Many search engines now have a separate site to search for images.

- The filename is the first element we can optimize, since it can be parsed for words. Rather than name an image of a rose **1.png**, we should call it **rose.png**.
- The judicious use of the alt attribute in the tag is another place where some textual description of the image can help your ranking.
- Finally, anchor text, like the text in URLs. If you have a link to the image somewhere on our site, you should use descriptive anchor text such as “full size image of a red rose,” rather than generic text “full size.”

Content

White Hat Technique

It seems odd that content is listed as an SEO technique, when content is what you are trying to make available in the first place.

- search engines tend to prefer pages that are updated regularly over those who are static
- If your website allows users to comment or otherwise write content on your site, you should consider allowing it.
- Entire industries have risen up out of the idea of having users generate content

Section 6 of 6

BLACK-HAT SEARCH ENGINE OPTIMIZATION

Black-Hat SEO

Do not use these techniques

Black-hat SEO techniques are popular because at one time they worked to increase a page's rank.

- these techniques are constantly evolving
- Google and other search engines may punish or ban your site from their results if you use black-hat techniques

Content Spamming

Black-Hat SEO

Content spamming is any technique that uses the content of a website to try and manipulate search engine results and include:

- Keyword Stuffing
- Hidden Content
- Paid Links
- Doorway Pages - inserting results for particular phrases with the purpose of sending visitors to a different page

Keyword Stuffing

Black-Hat SEO

Keyword stuffing is a technique whereby you purposely add keywords into the site in a most unnatural way with the intention of increasing the affiliation between certain key terms and your URL.

- As keywords are added throughout a web page, the content becomes diluted with them.
- Meaningful sentences are replaced with content written primarily for robots, not humans.
- Any technique where you find yourself writing for robots before humans, as a rule of thumb, is discouraged.

Hidden Content

Black-Hat SEO

Once people saw that keyword stuffing was effective, they took measures to stuff as many words as possible into their web pages.

Soon pages featured more words unrelated to their topic than actual content worth reading.

In response, rather than remove the unwieldy content, many chose to hide useless keywords by making them the same color as the background

This technique is detected and punished

Paid Links

Black-Hat SEO

Buying **paid links** is frowned upon by many search engines, since their intent is to discover good content by relying on referrals (in the form of backlinks).

Purchased advertisements on a site are not considered paid links so long as they are well identified as such, and are not hidden in the body of a page. Many link affiliated programs (like Google's own AdWords) do not impact PageRank because the advertisements are shown using JavaScript.

Doorway Pages

Black-Hat SEO

Doorway pages are pages written to be indexed by search engines and included in search results.

Doorway pages are normally crammed full of keywords, and effectively useless to real users of your site.

These doorway pages then link to your home page, which you are trying to boost in the search results

Link spam

Black-Hat SEO

Since links, and backlinks in particular, are so important to PageRank, and how search engines determine importance, there are a large number of bad SEO techniques related to links.

- Hidden Links
- Comment Spam
- Link Farms
- Link Pyramids
- Google Bombing

Hidden Links

Black-Hat SEO

Hidden links are as straightforward as hidden content.

With hidden links websites hide the color of the link to match the background, hoping that

- real users will not see the links.
- Search engines, will follow the links, thus manipulating the search engine without impacting the human reader.

Comment Spam

Black-Hat SEO

When you first launch a new website, going out to relevant blogs and posting a link is not a bad idea. After all you want people who read those blogs to potentially follow a link to your interesting site.

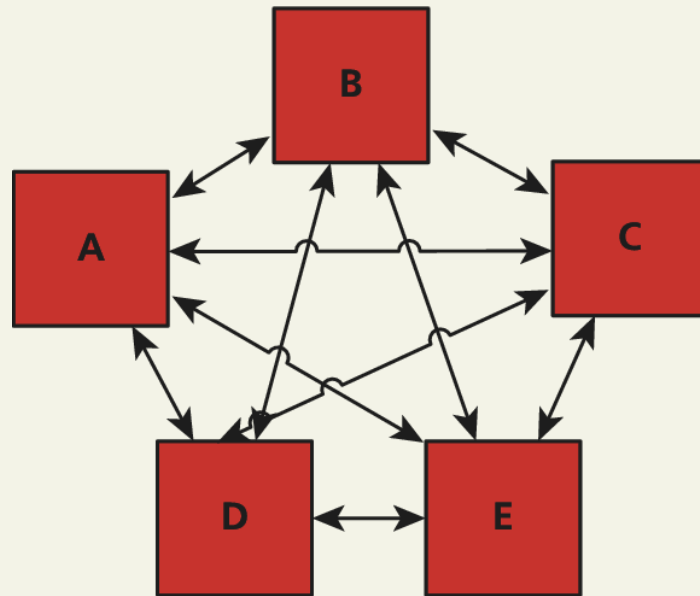
Since adding actual comments takes time, many spammers have automated the process and have bots that scour the web for comment sections, leaving poorly auto-written spam with backlinks to their sites.

If you have a comment section on your site, be sure to similarly secure it from such bots, or risk being flagged as a source of comment spam.

Link Farms

Black-Hat SEO

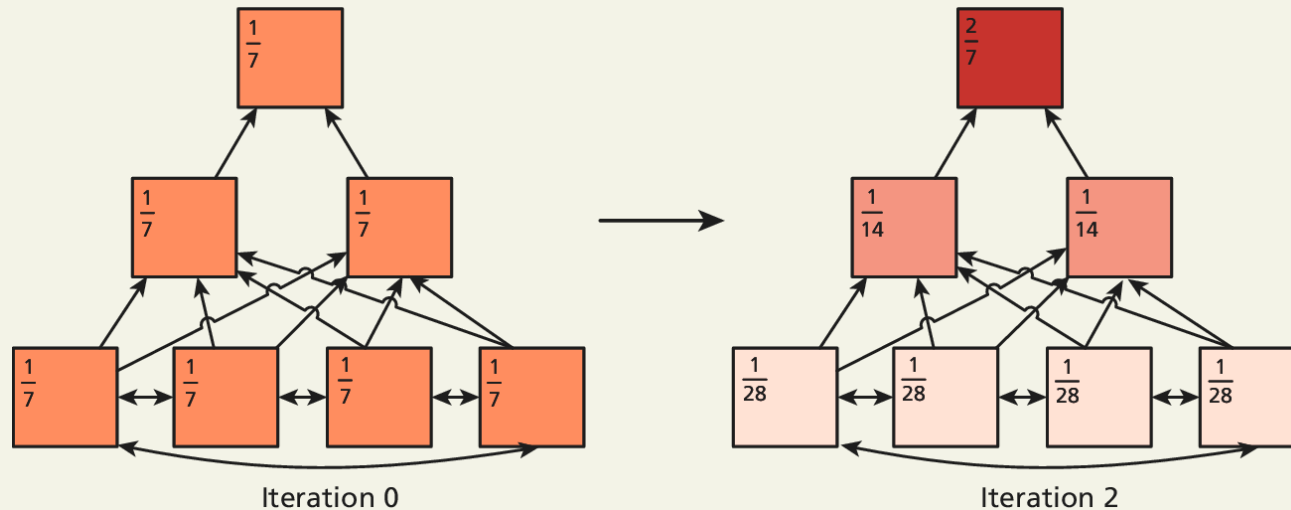
A **link farm** is a set of websites that all interlink each other with the intent of sharing any incoming PageRank to any one site with all the sites that are members of the link farm.



Link Pyramids

Black-Hat SEO

Link pyramids are similar to link farms in that there is a great deal of interlinking. Unlike a link farm, a pyramid has the intention of promoting one or two sites.



Google Bombing

Black-Hat SEO

Google bombing is the technique of using anchor text in links throughout the web to encourage the search engine to associate the anchor text with the destination website.

In 2006, webmasters began linking the anchor text “miserable failure” to the home page of then president George W. Bush. Soon, when anyone typed “miserable failure” into Google, the home page of the White House came up as the first result.

Other Spam Techniques

Black-Hat SEO

Although content and link spam are the prevalent black-hat techniques for manipulating search engine results, there are some techniques that defy simple classification.

Google Bowling

Cloaking

Duplicate content

Google Bowling

Black-Hat SEO

Google bowling is a particularly dirty and immoral technique since it requires masquerading as the site that you want to weaken (or remove) from the search engine

1. black-hat techniques are applied as though you were working on their behalf. This might include subscribing to link farms, keyword stuffing, commenting on blogs, and more
2. report the competitors' website to Google for all the black-hat techniques they employed!

Cloaking

Black-Hat SEO

Cloaking refers to the process of identifying crawler requests and serving them content different from regular users.

A simple script can redirect users if *googlebot* is the user-agent to a page, normally stuffed with keywords

Serving extra and fake content to requests with a known bot user-agent header can get you banned.

Duplicate Content

Black-Hat SEO

Stealing content to build a fake site has worked in the past, and is often used in conjunction with automated link farms or pyramids. Search engines are starting to check and punish sites that have substantially duplicated content.

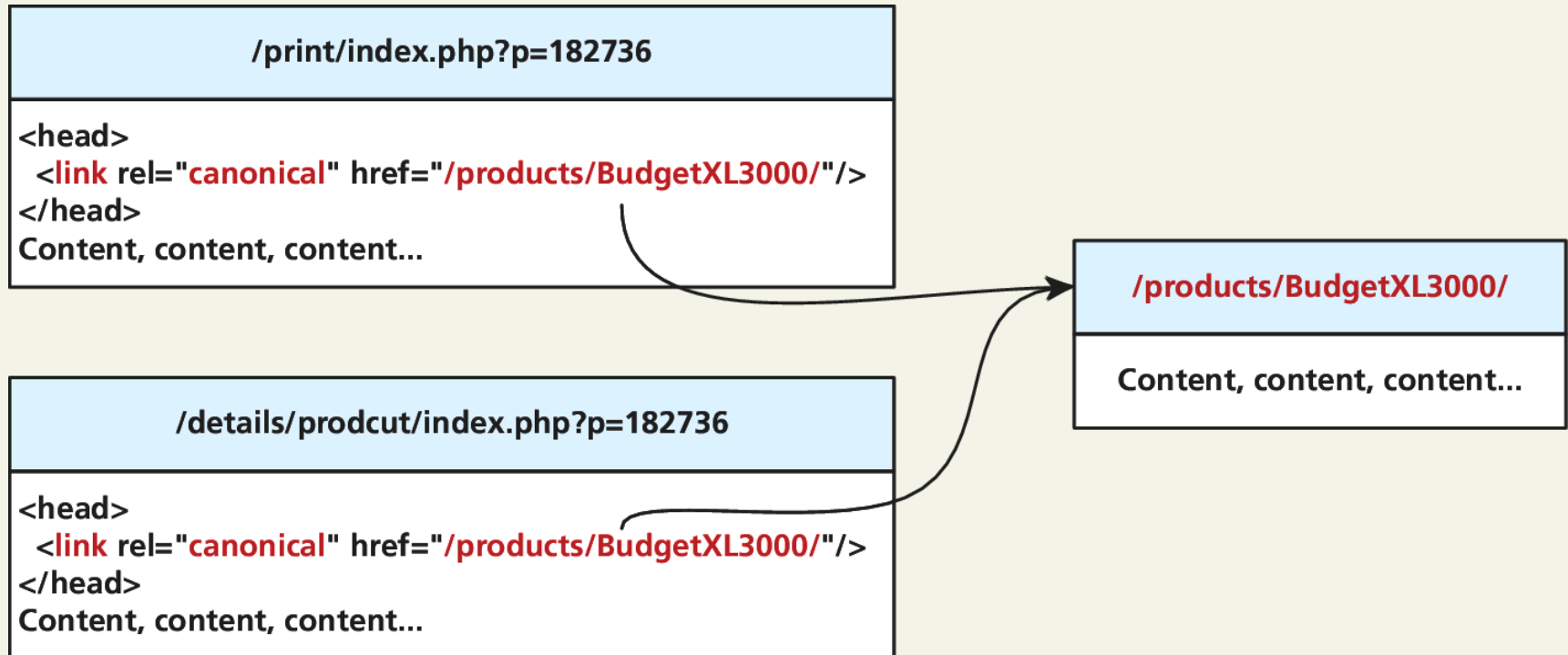
To attribute content to yourself use the rel=author attribute. Google has also introduced a concept called Google authorship through their Google+ network to attribute content to the originator.

Duplicate Content

Black-Hat SEO

Sometimes you have several versions of a page, for example, a display and print version.

To prevent being penalized, you can use the **canonical** tag in the head section of duplicate pages to affiliate them with a single canonical version to be indexed.



What You've Learned

1

The **History** and **Anatomy**
of Search Engines

2

Web **Crawlers** and Scrapers

3

Indexing and Reverse
Indexing

4

PageRank and Result
Order

5

White-Hat Search Engine
Optimization

6

Black Hat Search
Engine Optimization