ANALYSIS OF A SECOND HAND GOOGLE MINI SEARCH APPLIANCE

Stephen Larson
Slippery Rock University of Pennsylvania
Slippery Rock, PA
stephen.larson@sru.edu

ABSTRACT

Information and the technological advancements for which mankind develops with regards to its storage has increased tremendously over the past few decades. As the total amount of data stored rapidly increases in conjunction with the amount of widely available computer-driven devices being used, solutions are being developed to better harness this data (LaTulippe, 2011).

One of these solutions is commonly known as a search appliance. Search appliances have been used in e-discovery for several years. The Google Mini Search Appliance (Mini) has not only been used for e-discovery, but for indexing and searching internal documents. To accomplish these tasks, search appliances not only cache html versions of the documents, they contain metadata about the indexed documents, as well as metadata about search activity. This research analyzes the Mini to determine what forensically interesting artifacts exist on the device.

Keywords: forensics, digital forensics, Google Mini, search appliance

1. INTRODUCTION

Search appliances have a unique ability to crawl intranets and file servers and index the results–a function that is quite effective for e-discovery. Corporations implement search appliances not only for e-discovery, but for use in enterprise search to aid in finding of electronically stored information (McElhaney & Ghani, 2008). The legal community has considered the effectiveness of search appliances for e-discovery and adopted its use (Chan, 2009; Cuff, 2009). Likewise, the Google Mini Search Appliance (Mini) has also been used for e-discovery (Burgess & Metz, 2008).

The Mini is a 1-rack unit device and is considered an appliance instead of a server. The entire Mini system–hardware, software, and license for support–costs under $2,000 for 50,000 documents and $9,000 for 300,000 documents, plus a yearly support fee. It is designed to crawl your Web sites and file systems, index and cache the content, deliver search results, and enable users to produce reports that deliver valuable business data. It is not designed to search

more than 300,000 documents, crawl relational databases, or integrate with third party security software.

The Mini is touted to provide Google-quality enterprise search capabilities with no special tuning or coding. The Mini is offered with a perpetual license, including two years of support and hardware replacement coverage. It is a complete hardware and software appliance that was designed to run on intranets and internal company networks, requiring no specific expertise to configure. Indeed, "no Operating System administration is required, nor even possible" (Larrieu, 2009). It can crawl and search up to 300,000 documents, and provides support for user-level document security (Google, 2013a).

Information on the number of Minis sold or currently in use was not available. Unfortunately, due to the expanding breadth of its overall enterprise search offering, Google officially ceased production, sale and distribution of the Mini on July 31, 2012. The Mini will continue to function until the end of owner's current license agreement; the two year hardware warranty and technical support will be honored until the end of the two year period (Google, 2013b). The perpetual license allows owners to continue using the Mini, without support, until the owner retires the device.

The Mini's end-of-life suggests that companies will start upgrading to Google Site Search or Google Enterprise Search, and dispose of their Mini search appliance. A recent search on e-Bay returned eight Mini devices up for auction.

2. WHAT CAN BE FOUND ON A MINI?

Before examining the Mini, we decided to explore what could be found on it, what capabilities and limitations it has, and what reports we could expect to find. The information in this section can be found on Google's support Web site (Google, 2013c; Google, 2013e). Like its enterprise counterpart, the Mini can crawl and index over 200 different file formats, in several different types of locations. Conversely, the Mini uses the same search engine technology as the enterprise search appliance, and most of the information contained herein could be applied to a Google Enterprise Search appliance.

2.1 Search and Indexing Capabilities

The Mini can search and index internet sites, intranets, file systems, metadata, and document/download libraries. It supports over 220 different file formats, but is used mainly for HTML, PDF, and Microsoft Office file formats. With internal documents, the Mini provides a link to the document, so only users who have rights to view it normally will be able to access it, but can cache an html version of the document, which is accessible via the search interface. Any document available via HTTP and HTTPS protocols can be indexed. The Mini indexes XML as clear text. It can index up to 2.5 MB of clear text per XML

file. The Mini is designed to be able to index dynamic content in the same way that Google.com does.

The Mini can crawl and index UNIX and Windows file systems that are web-enabled or by using the SMB protocol. Metadata is fully searchable for Microsoft documents, HTML files, and PDF files. The Mini provides support for user-level document security using NTLM and HTTP Basic Authentication and Authorization. The Mini also can crawl secured content that requires LDAP authentication; subsequently the cached content is accessible via the search interface.

The Mini can crawl and index content in Lotus Domino. It can also crawl other e-mail servers if they are web-enabled. Database content is accessible to the Mini by web-enabling the database.

2.2 Provisos and Limitations

As with all search engines, the Mini has certain conditions and limitations:

- The Mini will honor robots.txt files and robot meta tags in the documents that it crawls. (Robots.txt is a plain text file that you create and put on your server to exclude search engine crawlers from accessing pages or directories on your site.)
- The Mini doesn't index the actual content of the video, graphics and audio files, though they will be included based on filename and metadata.
- The Mini can crawl and index files of up to 30 MB. Files that are larger than 30 MB are discarded without being indexed. If an HTML file is under 30 MB, the search appliance indexes the first 2.5 MB and discards the rest of the file. If a non-HTML file is under 30 MB, the search appliance converts the non-HTML file to HTML. If the converted content is less than 4,000,000 bytes, the search appliance indexes the first 2 MB of the HTML. The remainder of the file is discarded. If the converted content is more than 4,000,000 bytes, the document is not indexed, but the document and a link to the document appear in search results.
- Java Applets are not indexed, nor will the Mini crawl through URLs contained within JavaScript code.
- The Mini does not crawl and index content in MS SharePoint or other content management systems.
- The Mini is not able to integrate with single sign-on systems.

2.3 Reports and Logs

The Mini provides reports, logs and diagnostics, including diagnostics of crawling and indexing. All of these reports are accessible from the admin page.

The Mini's Crawl Diagnostics Reporting provides information about the current index and the URLs in the search index, including:

- Total Inflight URLs: The total number of URLs that have been identified but not yet crawled.
- Total Crawled URLs: The total number of URLs crawled at the time of viewing the page, including Locally Crawled URLs.
- Locally Crawled URLs: Pages directly fetched from the production index instead of from the actual Web site. This can be disabled with the "Recrawl all Pages" option on the Crawler Parameters screen.
- Retrieval Errors: URLs that could not be reached by the crawler because the server returned an error for them, possibly due to network problems.
- Excluded URLs: The URLs that were discovered, but dropped and not crawled at all. URLs are excluded by "Do Not Crawl" patterns and by robots.txt files.
- Crawling Rate: The current crawling rate, listed in pages per second.
- Total size of the stored documents: Total file size of the pages crawled.

The Mini has several summary reports which provide the following data:

- Total Results pages: The number of result pages seen by users for the report period. This includes both search results and non-search results, such as requests of cached pages. This value includes every result page viewed.
- Total Searches: The total number of search result pages seen by users. If a user performs a search and then selects "next" to see a second page, that counts as two searches.
- Distinct Searches: The number of times users submitted a specific search. Distinct Searches only include the first page where the user typed in a search but not subsequent pages for the same query.
- Number of Searches per Day.
- Average Number of Searches Per Hour.
- The Top 100 Keywords and number of Occurrences for each keyword.
- Top 100 Queries and number of Occurrences for each query.
- Average Result Sets per Query: The ratio of Total Searches to Distinct Searches equals your Average Result Sets per Query. This represents, on average, how many pages of search results a user views for each search he/she does.

The Mini's main log file is the Event Log, an audit trail of all system activity, including:

- Logins and logouts of users.
- Date and time of crawling (when the crawl was paused and resumed).

- Creations of collections and front ends.
- Serving index rollback time, if one occurred.
- Date and time of system password change.

2.4 Search Results

The search results contain metadata about the indexed documents, including a cached HTML copy, just as with Google Site Search or Google Enterprise Search. For this reason, companies can use search appliances like the Mini for data backup. While the Mini isn't as comprehensive as a full backup solution, it is probably less onerous than searching through backup server for vital documents during a temporary outage.

As noted in Section 2.1, the documents listed in the search results may not be accessible by individuals without the correct access permissions because the link provided is the link to the actual location of the documents. However, a cached HTML copy may also exist, without access limitations (see Appendix).

Because of these capabilities, there is a possibility for confidential documents, personally identifiable information, and corporate secrets to be saved in the cache areas.

3. ANALYSIS

A Mini was purchased from a used computer store on eBay. To gain insight on what to expect upon opening the Mini, we studied Clark's (2005) and Garrison's (2012) examinations of the Mini. The result of our initial inspection was similar to Clark's–the Mini had proprietary screws (see Figure 1) to inhibit opening the server.

Figure 1 Proprietary Screw on Google Mini Search Appliance (Clark, 2005)

Like Clark's, our Mini was 1-rack unit in size, the original equipment manufacturer was Gigabyte (motherboard and CPU), and it contained two

Pentium III-S processors running at 1.26 GHz. All of the drive bays had PATA interfaces and are handled directly by a Promise IDE RAID controller.

The memory consisted of 2 GB of PC133 SDRAM (4 x 512 MB sticks), which were branded as Dell memory but the chips were made by Micron.

The focus of this paper is the contents of the HDD. As the internals of the Mini were congruent with the findings of Clark (2005) and Garrison (2012), the details of the motherboard, ports, slots, power supply, etc., will not be discussed.

3.1 HDD Analysis

After enduring the frustration of removing 23 screws without being able to remove the HDD, and hesitant to drill out the remaining screws, we contacted the original manufacturer for instructions on how to remove the HDD and were told that the Google Mini box is a custom made box and to contact Google for instructions. Requests for instructions were not possible as this is a second-hand search appliance whose support period had expired and as such no support is offered. The used computer store from which this Mini was purchased was contacted for more information; they replied that the HDD had been wiped to DoD standards as per company policy and in compliance with the End User License Agreement (Google, 2013f).

Due to the seeming impossibility of extracting the HDD, the decision was made to duplicate the HDD in place. The HDD was disconnected from the motherboard and power supply and connected to a Tableau Forensic Duplicator.

The single HDD was a 120 GB Seagate Barracuda 7200.7 model number ST3120022A. Relevant S.M.A.R.T. (Self-Monitoring, Analysis and Reporting Technology) information is as follows:

Firmware revision 3.06
Capacity in sectors: 234,441,648 (120.0 GB)
HPA in use: No
DCO in use: No
ATA Security in use: No
Cable/Interface type: IDE
ATA PIO mode: PIO 4
ATA DMA mode: UDMA 5
Peak power:
+5V : 5.02 V 1.12 A
+12V: 12.5 V 1.87 A
Spin Up Time: 390
Power cycle count: 147
Power off Retract Count: 48

Power-on hours: 1074
Maximum temperature: 36C
Write error count: 200

Upon successful duplication, the HDD was forensically examined. Contrary to what the seller stated, the HDD had not been wiped and contained three Linux EXT 2 volumes, all MBR partitions, block size = 4k:

1) Linux Ext Volume 1 (~1 Gb): contained the Linux boot and OS – Red Hat Linux release 6.2 (Zoot); starting at sector 63.
2) Linux Ext Volume 2 (~110 Gb): contained Google Mini search programs, configuration information, log files, etc.; starting at sector 2056320.
3) Linux Swap Volume 1 (~517 Mb); starting at sector 3116610.
4) Unpartitioned space (~2.5 Mb); starting at sector 234436545.

There were no hidden partitions found on the HDD.

4. INTERESTING DATA

This section includes information (or lack thereof) on the location of data that has been considered to be of forensic interest.

4.1 Linux Ext2 Volume 1

The OS boot volume contained what one would expect to find on a computer running Red Hat Linux release 6.2. There appeared to be nothing special about the boot sector code. Settings reflected the previous owner's information, including:

The /etc/localtime showed the time zone set to MST/MDT.

The /etc/dhcpd.conf file contents:

```
default-lease-time 600; max-lease-time 7200; option subnet-mask
255.255.255.0; option broadcast-address 192.168.255.255; option routers
192.168.255.254; subnet 192.168.255.0 netmask 255.255.255.0 {range
192.168.255.254; }.
```

The /root/machine_install_log shows the Google Enterprise search software was installed on 4 May 2005.

The /etc/alias file contained one non-system name, "marc", who was to receive mail for root.

/etc/hosts file contents:

```
127.0.0.1            localhost
192.168.255.254      install_laptop
0.0.0.0                    localhost
216.239.43.1         ent1    ent1.ent.google.com.
216.239.43.2         ent2    ent2.ent.google.com.
…
216.239.43.127       ent127  ent127.ent.google.com
```

/etc/lilo.conf file contents:

```
restricted password=cfnt5FA7 boot=/dev/hdamap=/boot/map
install=/boot/boot.b prompt timeout=50
root=/dev/hda1
image=/boot/vmlinuz-2.2.19-ent      label=2.2.19-ent
image=/boot/vmlinuz-2.2.19-ow       label=2.2.19-ow
image=/boot/memtest.bin       label=memtestx86
```

The network and IP settings were found to be as follows:

Basic Network Settings:

```
IP address: 10.0.0.2
Subnet mask: 255.255.255.0
Default Gateway: 10.0.0.1
```

DNS Settings:

```
DNS Server: 10.0.0.1
DNS Suffix (search path): /etc/lib
```

These network settings suggest that this Mini was configured for internal use on an intranet and not for an internet-facing audience.

Mail Server Settings:

```
SMTP: smtp.google.com
Sender of outgoing mail: nowhere@somewhere.com
```

Time Settings:

```
Local Time Zone: Mountain
NTP Servers: ntp.google.com
```

Administrator account:

```
Username: admin
Email address: somewhere@nowhere.com
```

In combination with the low power cycle count and power-on hours from the S.M.A.R.T. information, these settings suggest this search appliance was never put into actual production use, but was perhaps only a test unit.

4.2 Linux Ext2 Volume 2

The Ext Volume 2 with the Mini search programs and configuration files naturally contained the most useful forensic information. The root directory contained the following folders:

/3.4.14
/distribution
/logs
/lost+found
/spelling-data
/support_records
/tmp
/versionmanager

The folder named 3/4/14 contained the version of Google Enterprise search, and the following subfolders: data, local, querycache, spelling, tmp folders. These folders and subfolders contained configuration and data about the Mini, including the following:

- The URL of the organization
- The root URL from which crawling will start
- The patterns of text for URLs that are excluded from the crawl
- Excluded filetype extensions
- Whether the Mini is currently crawling and populating the staging index or the production index
- The current crawl status, including the number of each type of crawl instance (numbers found in parentheses):
    - total InFlight URLs (0)
    - total crawled URLs (26,869)
    - locally crawled URLs (320)
    - retrieval errors (766)
    - Excluded URLs (365)
- Queries, searches and search results, cached documents in a proprietary format
- All the different words found during the crawling, including misspelled words – crawling dictionary
- Apache and tomcat pages and information:
    - Web pages that appeared to be administrative pages for configuring the search collections and parameters
    - image files

- Log files from searches, including search date/time, search terms, originating user/IP address, results returned
- Different spellings of words found during the crawl
- Addedurls files (URLs found during crawls)
- Status of index builds
- Python scripts and web pages for the administrative pages
- License information for the Mini
- Appliance ID and License ID
- System event logs
- Various reports, including:
    - Searches per day and average searches per hour
    - Top 100 keywords searched
    - Top 100 queries

Additionally, there were many deleted files which were retrievable via hex editor and/or carving.

These folders and files included forensically interesting information such as usernames, cc:mail information, organization information, files, employee names and personally identifying information, customer information, building and plant information and images, and even names and email addresses of Google employees. As shown in the Appendix, this particular Mini contained approximated 881 Mb of stored documents in its cache. Unfortunately, the files were saved in a proprietary format. The files were contained in 15,183 inodes.

Several locations contained information with potentially useful forensic information:

> \3.4.14\spelling contained current and expired spelling "dictionaries" which could be used to determine names or terms with uncommon spellings that had been searched.
>
> \3.4.14\querycache contained cached data about searches and the results of those searches, in a proprietary format (not in clear text).
>
> \3.4.14\data contained data and information about expired and current indexes, URLs, among other search-related items. File names containing the word "bigfile" were found, suggesting that the Mini uses the BigFile database (but this could not be confirmed).

This Mini also had a license valid until 2 Dec 9009 and could index up to 100,000 pages in one collection.

The log files were interesting in that they contained specific searches, among other things. For example, the contents of the query log file named weblog.from_ent1.port8888.starts20051219.log contained the following:

```
10.10.1.210 - - [19/Dec/2005:11:13:24 -0700] "GET
/search?ie=&q=joe+janson&site=testcollection&output=xml_no_dtd&clie
nt=testcollection&access=p&lr=&ip=10.10.1.210&proxystylesheet=testcol
lection&oe= HTTP/1.1" 200 763 0 1.05
```

The log format is an extension of the Common Log Format. Each line shows the host IP address of the requester (in this case 10.10.1.210), the date/time of the request (19/Dec/2005 at 11:13:24am), the requested "text" in double quotes (which includes the collection to which the request is made), followed by a three-digit status code (in this case, 200), the number of bytes returned to the requestor (763 bytes), the number of search results (in this case 0), and the total time in seconds that it took to fulfill the request (1.05 seconds).

4.3 Linux Volume 3

This volume was the swap file, and had an unrecognized file system. The volume was 542,868,480 bytes, and contained items that were written to the swap file by the OS, such as the OS, the Web server (apache), search engine configuration settings, recent search results, etc. The information in clear text was viewable via hex editor.

5. DISCUSSION / CONCLUSION

As seen above, forensically interesting data could be found. The cached documents contained company and personnel information, including birthdates, contact information, and schedules, and customer information. The cached documents are most easily accessible via the search function of the search appliance.

In particular, the log files contain several useful artifacts:

- Crawled and not crawled URLs might be useful in e-discovery requests.
- All words/terms found while crawling the site's document repositories, including misspelled words.
- Which URLs were included in, or excluded from, the crawl.
- A history of all searches performed by users. Coupled with the PC's logs and DHCP logs (when applicable), this information would be useful when proving if an individual searched for confidential or private information using the company's search appliance.

During the course of any forensic examination, there is always a chance of discovering interesting information that might prove embarrassing if made public, and this time was no exception. A file that appears to be a Google support file named

3/4/14/local/config/crawls/testcollection/bypass_robots.testcollection contained this text:

```
# Use only in case of absolute urgency for 8-way customers – this should
# never be exposed in the UI. First berate the customer, and ask them to
# fix their robots.txt/ suggest proxy servers before agreeing to this
# feature -- support will have to login to "maintain" this file
```

With the end-of-life of the Mini and more customers moving to Google site search or an enterprise search appliance, we expect more Mini search appliances on the second-hand market.

Additionally, beyond using the Mini or other search appliance just as a tool for e-discovery, it is worth taking the time to examine the metadata contained on search appliances during an investigation or as part of an e-discovery effort, particularly an internal investigation.

Appendix 1 contains the results of an examination of the Mini after booting it up with the duplicated HDD.

REFERENCES

Burgess, E., & Metz, E. (2008). Applying Google Mini search appliance for document discoverability. Online, *32*(4), 25-27.

Chan, A. (2009, July). Google to the (E-Discovery) rescue? Retrieved January 11, 2013, from eDiscovery: http://ediscovery.quarles.com/2009/07/articles/information-technology/google-to-the-ediscovery-rescue/

Clark, J. (2005). AnandTech Search goes Google. Retrieved January 11, 2013, from anandtech.com: http://www.anandtech.com/show/1781/3

Cuff, J. (2009). Key trends and developments of rights information management systems–An interview with Jim Cuff of Iron Mountain Digital. *Journal of Digital Asset Management*, *5*(2), 98-110.

Garrison, J. (2012, December 11). Google Mini Search Appliance teardown. Retrieved July 8, 2013, from http://1n73r.net/2012/12/11/google-mini-search-appliance-teardown/

Google. (2013a). Google Mini help. Retrieved January 11, 2013, from Google Web Site: http://support.google.com/mini/?hl=en#topic=219

Google. (2013b). Google Mini: Information. Retrieved January 11, 2013, from Google Web site: http://lp.google-mkto.com/NORTHAMSearchLCSMiniEndofLife_GoogleMiniFAQs.html

Google. (2013c). Google Mini report overview. Retrieved January 11, 2013, from Google Web site: http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.ie/en/ie/enterprise/mini/lbrary/MiniReports.pdf

Google. (2013d). First-time startup of a Google Search Appliance. Retrieved January 15, 2013, from Google Web site: https://developers.google.com/searchappliance/documentation/50/installation/InstallationGuide#FirstTime

Google. (2013e). Google Mini help center. Retrieved June 30, 2013, from Google Web site: https://developers.google.com/search-appliance/documentation/50/help_mini/home

Google. (2013f). Google Mini license agreement v3.0. Retrieved July 8, 2013, from Google Web site: http://1n73r.net/wp-content/uploads/2012/12/google-mini-eula.pdf

Larrieu, T. (2009). Crawling the control system. No. JLAB-ACO-09-1072; DOE/OR/23177-1007. Newport News, VA: Thomas Jefferson National Accelerator Facility.

LaTulippe, T. (2011). Working inside the box: An example of Google desktop search in a forensic examination. *Journal of Digital Forensics, Security and Law, 6*(4), 11-18.

McElhaney, S., & Ghani, S. (2008). *Enterprise search and automated testing. Governance, Risk, and Compliance Handbook: Technology, Finance, Environmental, and International Guidance and Best Practices*, 267.

APPENDIX: PLAYING WITH THE MINI

After finishing the forensic analysis, we decided to play around with the Mini. Before booting up the Mini with the duplicated HDD, we searched on Google's Web site for instructions on configuration and setup (Google, 2013d). The Mini was booted using the duplicated HDD with no network connectivity. The unit booted normally. Google's Web site states "The search appliance assigns the IP address 192.168.255.254 and subnet mask 255.255.255.0 to the computer connected to the search appliance." Hence a laptop was connected to the unit with a crossover Ethernet cable to the admin Ethernet port.

The configuration and setup instructions direct admins to browse to http://192.168.255.1:1111/. Upon browsing to this address, the "Google Search Appliance Network Installation" page appeared. As seen in Figure 2, the settings are as follows:

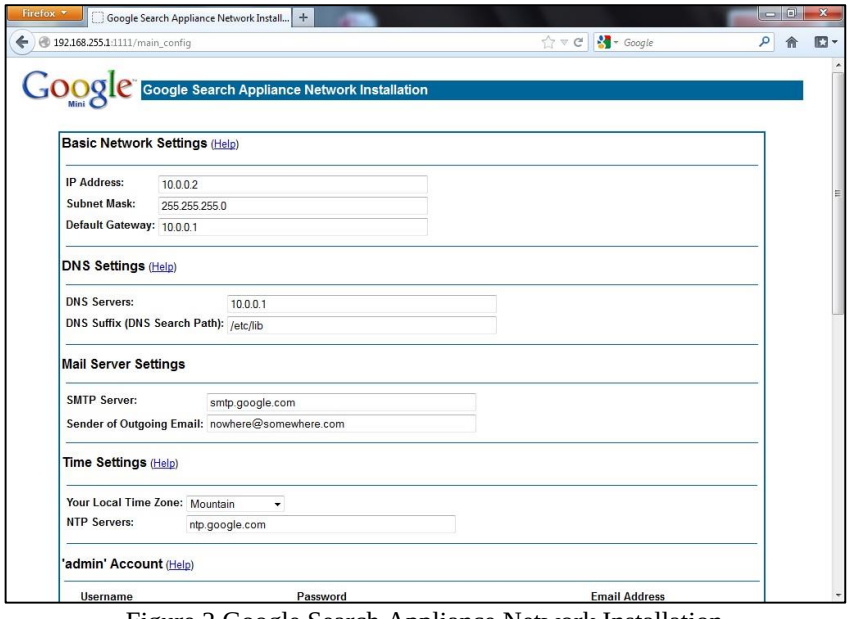

Figure 2 Google Search Appliance Network Installation

At this point, we changed the IP address on the laptop to 10.0.0.3 / 255.255.255.0 and browsed to https://10.0.0.2:8443 and reached the admin console log in page (see Figure 3) where we input the username and password given by the used computer store from which we bought the Mini.

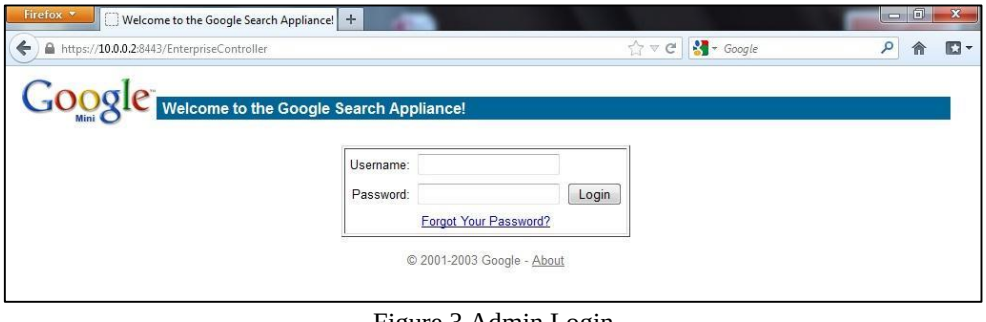

Figure 3 Admin Login

After logging in, the "Main" search configuration page opened (see Figure 4). As noted in the HDD Analysis section, a collection called "testcollection" was found.

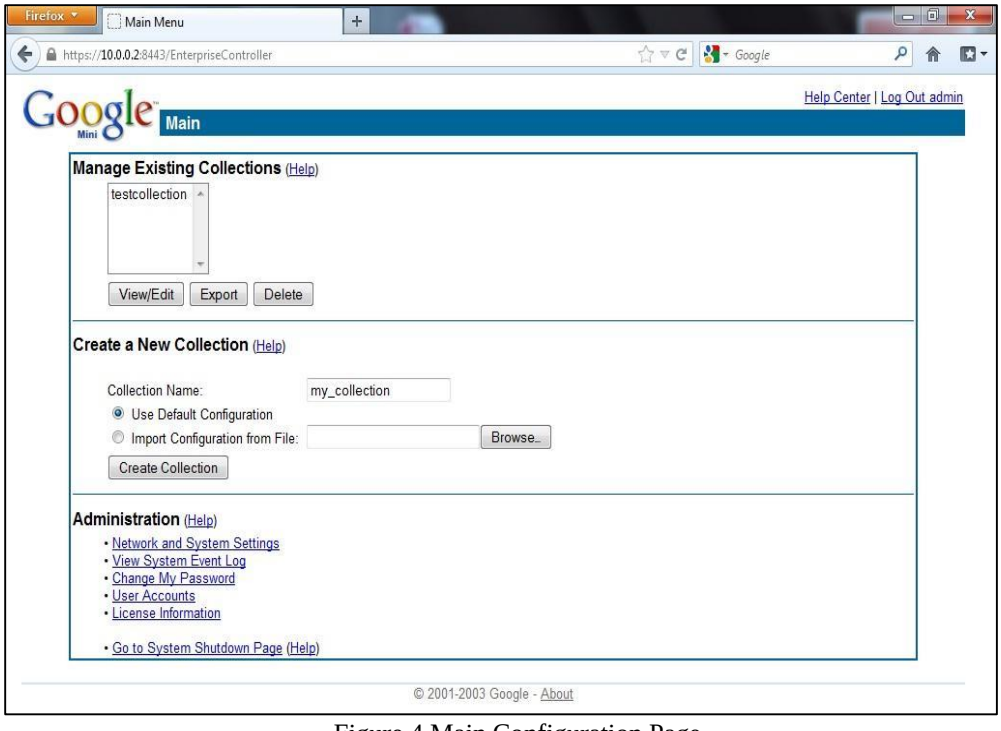

Figure 4 Main Configuration Page

Clicking the View/Edit button and brought up the settings for the "testcollection" collection (see Figure 5).

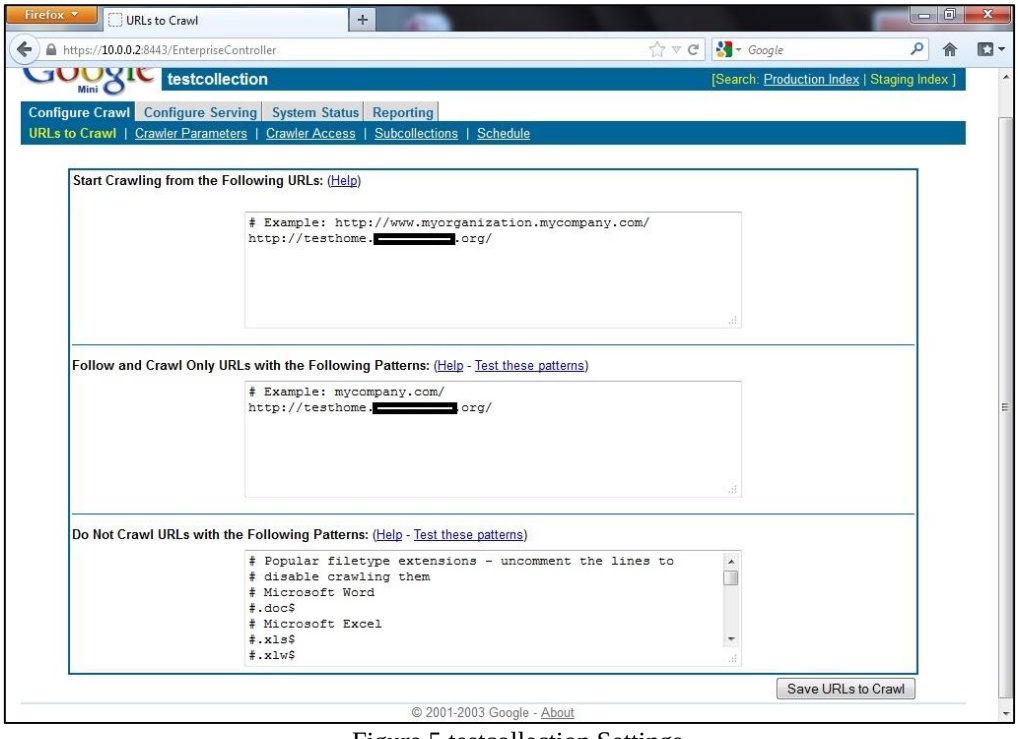

Figure 5 testcollection Settings

In the "Start Crawling from the Following URLs" box we found the org's name "xxxx.org". A quick search on the internet found this to be the domain of wholesale electric power supplier owned by the 44 electric cooperatives that it serves.

We then checked the Mini to determine if it could still serve up search results. To do so, we confirmed our laptop's IP address to be on the same subnet as the Mini's subnet (10.0.0.3 / 255.255.255.0), and browsed to https://10.0.0.2. A simple search page was presented, as shown in Figure 6.

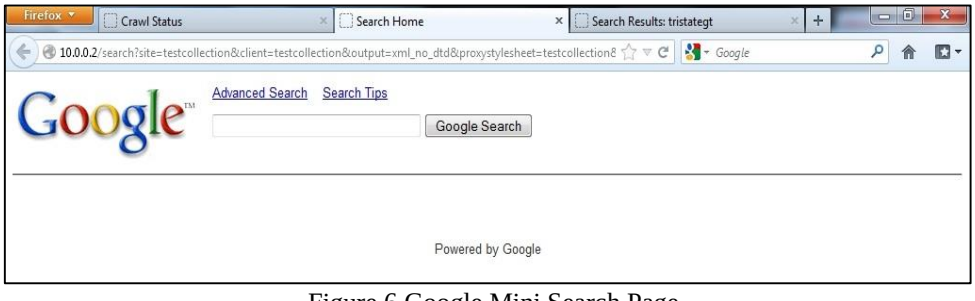

Figure 6 Google Mini Search Page

Searching various keywords return expected results, and we confirmed that thousands of documents were indeed cached on the Mini. A search for the

organization's name returned over 8,500 results. Additionally, a search for "birthday" returned a document with employee names and birthdates. Happy Birthday Sue!