Extending Yioop abilities to search the Invisible Web

Advisor/Committee Members

Dr. Chris Pollett

Dr. Sami Khuri

Prof. Frank Butt

By:

Tannmayee Potluri

Agenda

- Motivation
- Project Goal
- Background
- Yioop!
- Invisible Web Resources
- Modifications to Yioop!
- Tests and Results
- Conclusion
- References

Motivation

- A plethora of information is hidden in the Web.

- Not crawled and indexed by traditional search engines.

- The information hidden can be of great use and importance to the user.

- User might want to index different formats of data.

- It is useful to be able to crawl and index such data.

Project Goal

- Yioop! is an open source search engine written in PHP.
- The goal is to extend Yioop! to index few resources of the Invisible Web.
- To make Yioop! crawl and index log files.
- To make Yioop! crawl and index database records based on the query entered by user.
- Design user interface for customized crawling and indexing of log files and databases.
- To add code to make Yioop! deal short links appropriately.

Background

- Invisible Web : part of the web that is not indexed and is not a part of the surface web.

- It is a lot larger in magnitude than the surface web.

- The invisible web may constitute many resources:
  - Database content
  - Content in specific file formats like log files
  - JavaScript links
  - Password protected sites
  - Short links

- It is an added feature to have search engines index a part of the Invisible Web.

Existing tools

- There are few existing tools in the field of log files and databases.

- AWStats, Webalizer and Analog are the log file analyzers available that provide statistical information about log files.

- The data is provided in the form of graphical web pages.

- Sphinx is an open source search engine developed solely for indexing database content.

Yioop!

- Yioop! is an open source search engine written in PHP by Dr.Chris Pollett [1].

- Yioop! can be used both as a traditional search engine or it can be used for personal crawling.

- It can be modified to act as a search engine, only for a predefined set of domains or URLs [1].

- Yioop is licensed under GPLv3 and SeekQuarry is the parent site for Yioop [17].

Yioop! (Cont…)

➢Yioop has been designed in order to:

- ease the usage of personal crawls
- make it easier to crawl archives
- perform archive crawls.

➢ Yioop Requirements

- A web server (e.g. Apache)
- PHP 5.3 or higher
- PHP multi-curl library (for download of web pages)

➢ Xampp software has all the three necessary requirements of Yioop available in it.

Archive Crawling in Yioop!

- Archive crawling is an existing feature in Yioop.

- It is used to perform crawling of the previous crawls performed in Yioop or crawl one of the web archive file formats.

- These formats include like Arc, MediaWiki XML, ODP RDF.

- In order to crawl each of these file formats, separate archive bundle iterators have been written in Yioop.

- In order to perform an archive crawl in Yioop, one has to create a folder "archives" in the cache folder of the WORK_DIRECTORY of Yioop.

Archive Crawling in Yioop! (Cont...)

- In the archives folder, another folder should be created to store all the files of a particular format that are to be crawled.

- As an example, we will see how to create a folder to crawl log files that is implemented in this project.

- In order to perform a crawl on the log files, one needs to create a folder like my_log_files in the WORK_DIRECTORY/cache/archives folder.

- Each of those folders in archives folder should contain a file named arc_description.ini.

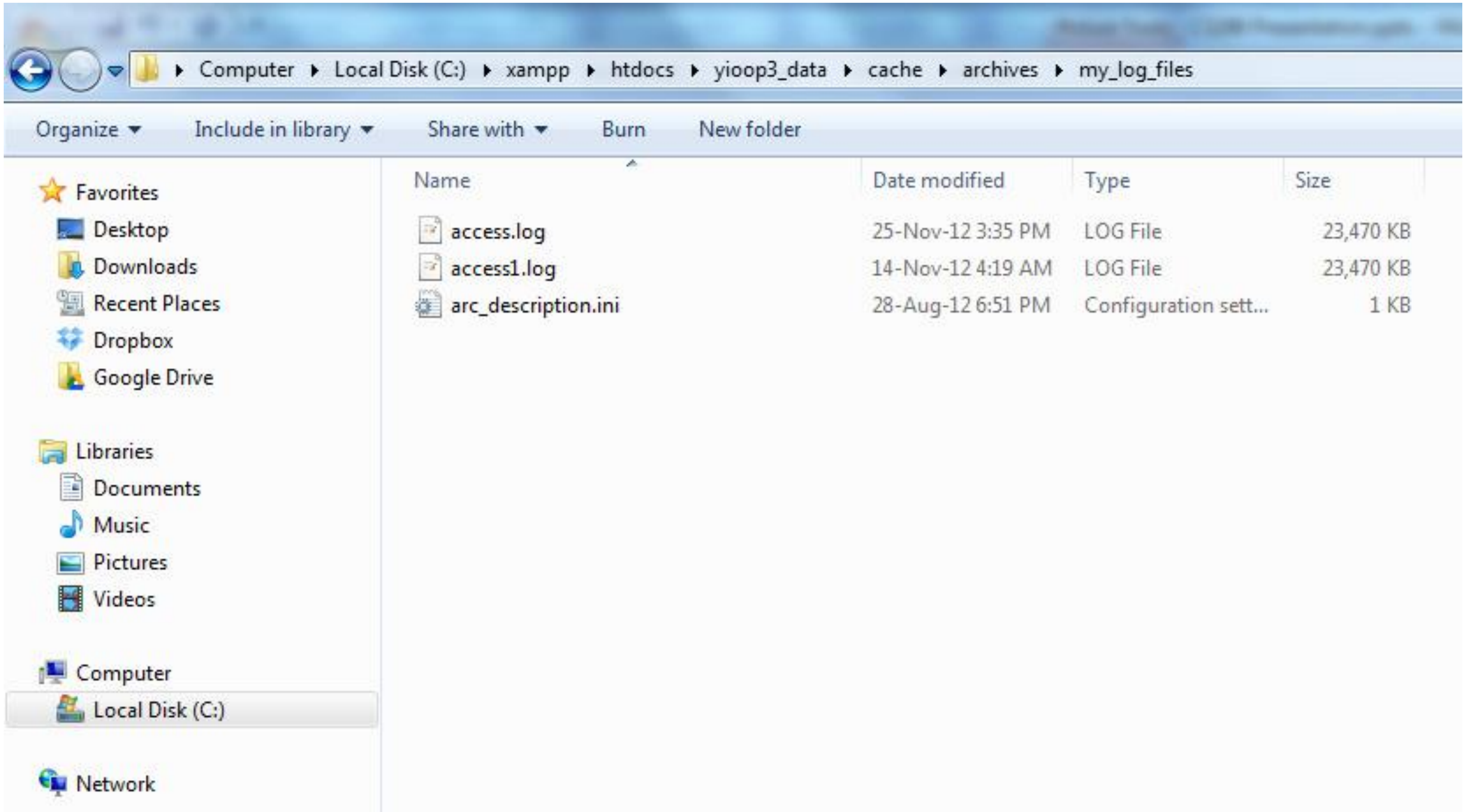
Archive Crawling in Yioop! (Cont…)

- The arc_description.ini file is a text file containing text like,

```
arc_type = 'LogArchiveBundle';
description = 'Log Files';
```

- The arc_type should be the “*filetype*ArchiveBundle”.
- The “*filetype*” changes for each of the file types.
- The description field can be anything the user wishes to see in the user interface.
- Once this folder is created ARCFILE::Log Files is seen as one of the options to recrawl.

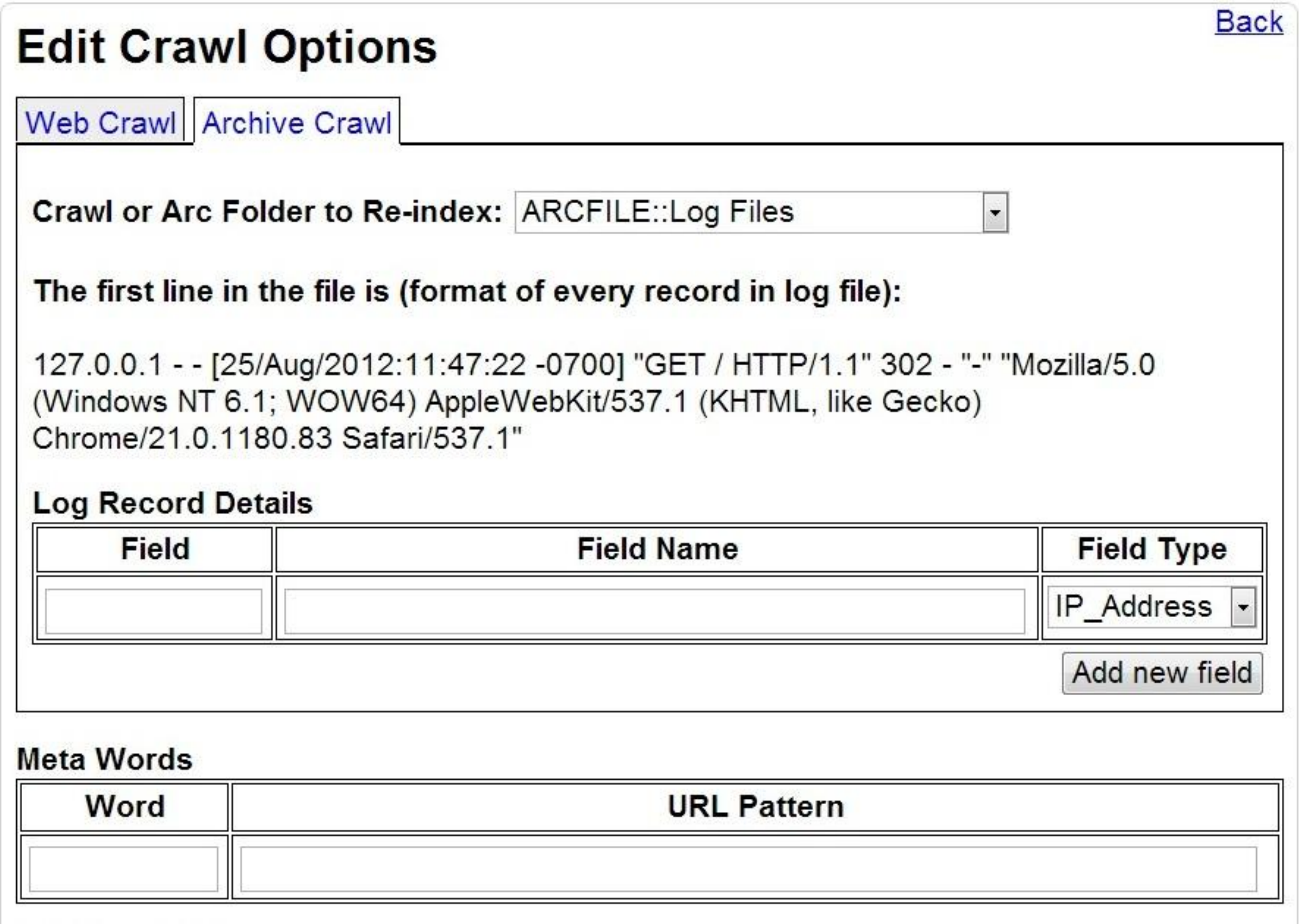
Archive Crawling in Yioop! (Cont...)



Name	Date modified	Type	Size
access.log	25-Nov-12 3:35 PM	LOG File	23,470 KB
access1.log	14-Nov-12 4:19 AM	LOG File	23,470 KB
arc_description.ini	28-Aug-12 6:51 PM	Configuration sett...	1 KB

Figure: Folder to store log files to be crawled

Archive Crawling in Yioop! (Cont…)



Edit Crawl Options

Crawl or Arc Folder to Re-index: ARCFILE::Log Files

The first line in the file is (format of every record in log file):

```
127.0.0.1 - - [25/Aug/2012:11:47:22 -0700] "GET / HTTP/1.1" 302 - "-" "Mozilla/5.0
(Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko)
Chrome/21.0.1180.83 Safari/537.1"
```

Log Record Details

Field	Field Name	Field Type
		IP_Address

Meta Words

Word	URL Pattern

Figure: Recrawl Options Interface for Archive Crawling

Yioop! Architecture

➢Yioop is designed using the web-based Model-View Controller framework [1].

➢The two main programs of Yioop are in the bin folder, namely, the fetcher and the queue_server.

➢ The queue_server program maintains a queue of URLs that are going to be scheduled.

➢The fetcher program downloads batches of URLs provided by the queue_server.

Yioop! Architecture (Cont…)

- The lib folder has the archive_bundle_iterators folder which contains all the archive bundle iterators needed to crawl and index different file formats in Yioop.

- An archive bundle iterator iterates over a particular format of files.

- The controllers folder contains all the controllers used by the Yioop search engine.

- The elements folder in Yioop contains the elements responsible for providing a view of the user interface.

- The files in these folders have been modified in this project.

Invisible Web Resources

➢The invisible web may constitute many resources like:

- Shortened Links
- Database content
- Any content in specific file formats
- JavaScript links
- Password protected sites

➢Three of these resources have been dealt in this project, namely,

- Log files
- Databases
- Shortened links

Log files

- Log files are one of those file formats that are not crawled and indexed by traditional search engines.

- Log files are the files to which a computer system writes a record of its activities [9].

- Log files are generally automatically created and maintained by a server containing the information of the activity performed by it [9].

- An example of a log file is an Apache access log file which maintains a history of page requests.

Log files (Cont…)

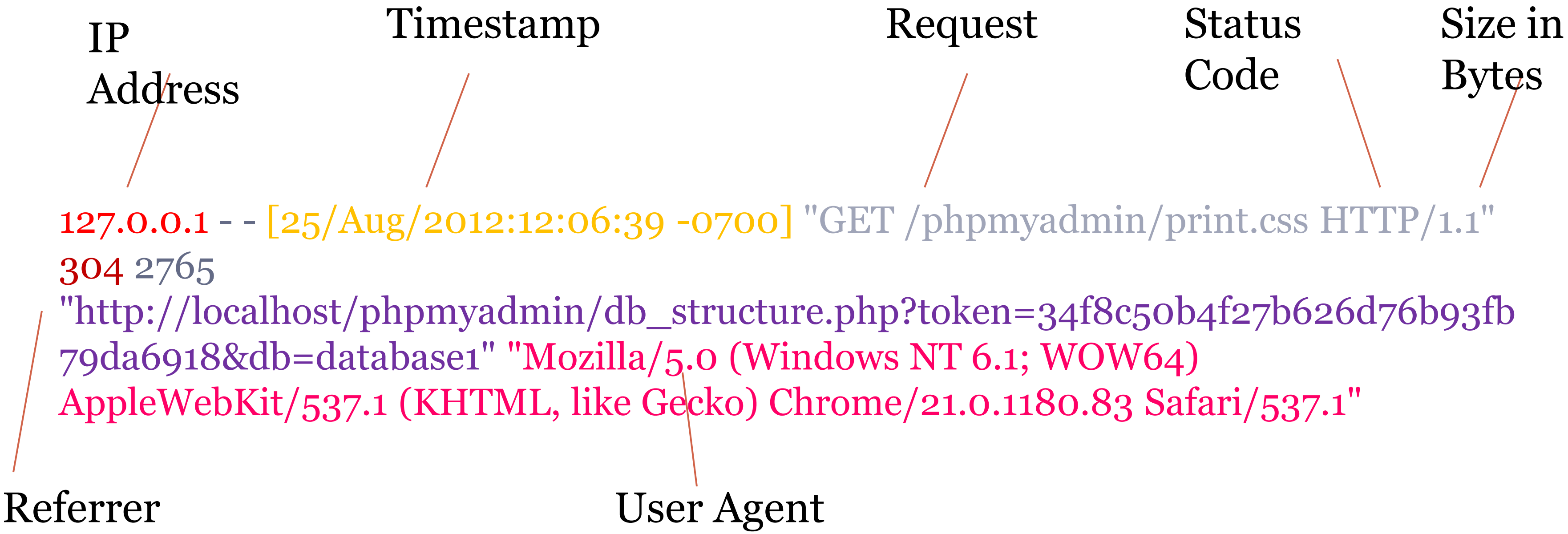
➢A common log file format is defined by W3C which is used by most of the servers to generate log files.

➢ The general predefined fields that are present in a log file are:

- IP Address
- Timestamp
- Request
- Status Code
- Size in Bytes
- Referrer
- User Agent
- - : Information not returned by the server

➢There can be few other formats for log files too.

Common Log file format



IP Address

Timestamp

Request

Status Code

Size in Bytes

```
127.0.0.1 - - [25/Aug/2012:12:06:39 -0700] "GET /phpmyadmin/print.css HTTP/1.1"
304 2765
"http://localhost/phpmyadmin/db_structure.php?token=34f8c50b4f27b626d76b93fb
79da6918&db=database1" "Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.83 Safari/537.1"
```

Referrer

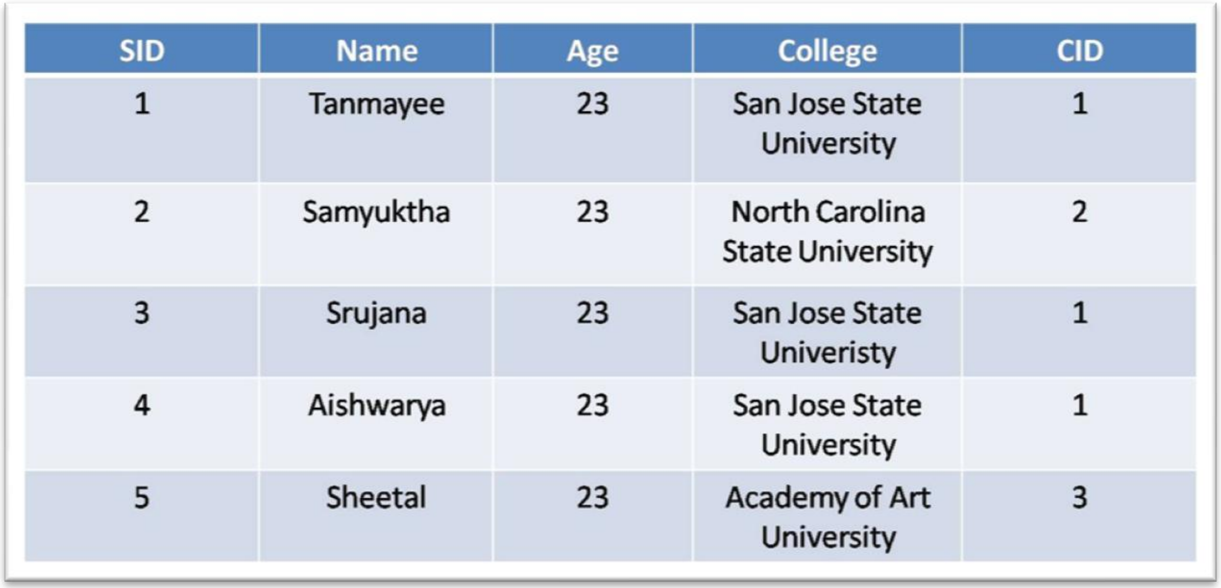
User Agent

Databases

- Database files are a collection of similar kind of records[12].

- A database can contain any number of tables and tables can contain any number of records.

- The database content is one kinds of deep web content that is not crawled by traditional search engines.



SID	Name	Age	College	CID
1	Tanmayee	23	San Jose State University	1
2	Samyuktha	23	North Carolina State University	2
3	Srujana	23	San Jose State Univeristy	1
4	Aishwarya	23	San Jose State University	1
5	Sheetal	23	Academy of Art University	3

URL Shortening Services

- URL shortening is a technique on the Web in which a URL is made short in length and still redirected to the required page.

- The rank or weight that should be given to the original URL is given to the short link which makes them a source of dark content on the Web.

- For example, the URL http://www.yahoo.com is shortened to bitly.com/4bYAV2.

- There are many URL shortening services of which bit.ly and tinyurl are few of them.

- The underlying address can be disguised using URL shortening services [14].

Modifications to Yioop!

➢Yioop! has been modified to add:

- A Log archive bundle iterator for crawling and indexing log files.

- A Database archive bundle iterator for crawling and indexing databases.

- User Interface for the users to input the necessary details for crawling log files and databases.

- Code to improve the search results for short links.

Modifications to Yioop! (Cont...)

In order to add indexing of log files and databases in Yioop, two tasks needed to be performed:

- Build a user interface for the users to input details for crawling
- Create an archive bundle iterator for each of the features.

In order to implement this user interface, the following files in Yioop had to be modified.

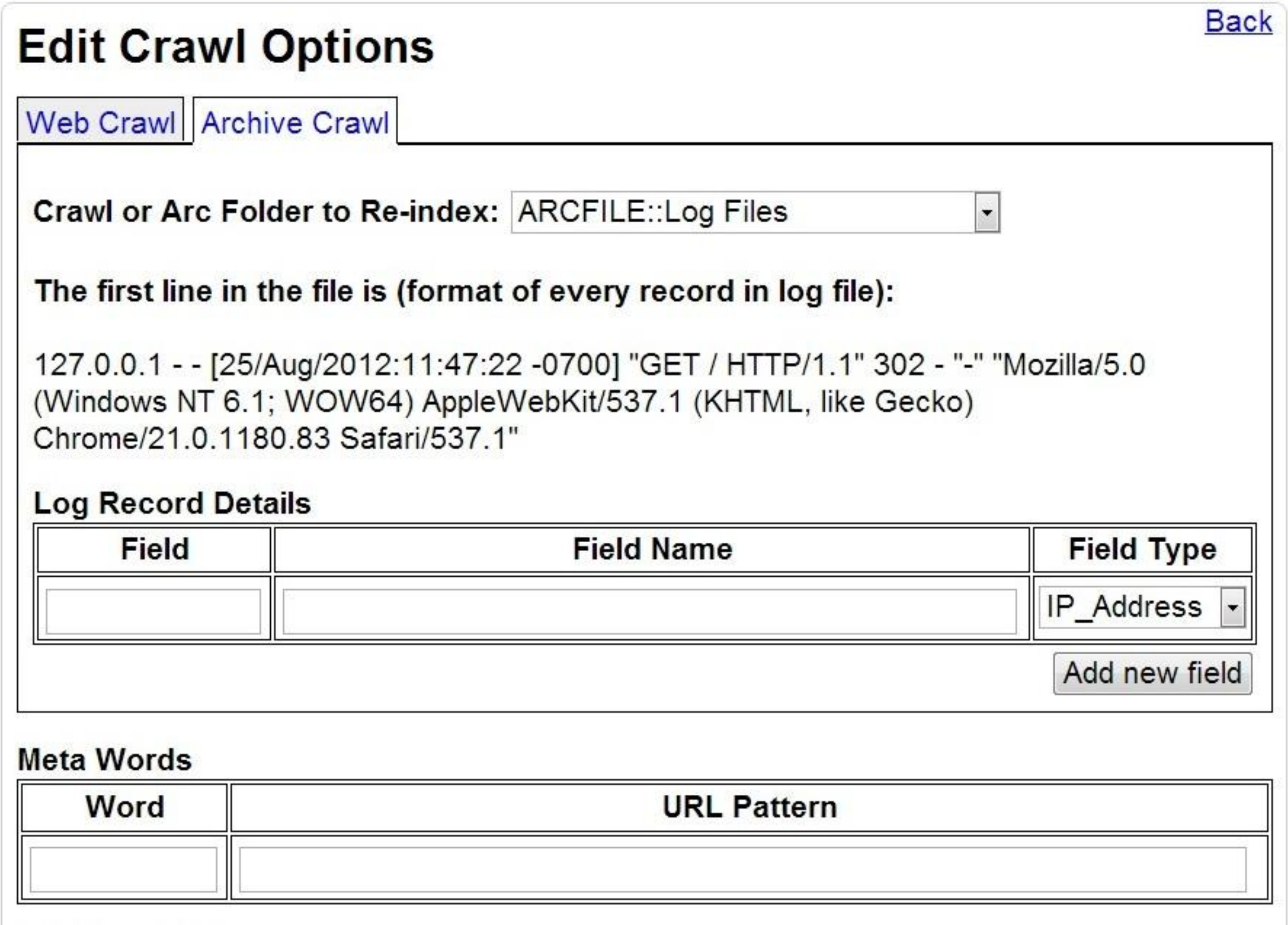
- ➢ **admin_controller.php**

  - The code to store the details of the log records entered by the user in the web interface of Yioop is written here.

Modifications to Yioop! (Cont…)

- **crawloptions_element.php**
  - The UI that needs to be displayed when log files is selected as an option is created in this file.
  - It retrieves the details of inputs given by the user from the admin_controller.php program and displays the options.
  - It also has the code to save the options entered by the user.
- **search.css**
  - The code is added here to specify the styles for the log records table shown in the web interface to the user.

Modifications to Yioop! (Cont…)



Back

Edit Crawl Options

Web Crawl | Archive Crawl

Crawl or Arc Folder to Re-index: ARCFILE::Log Files

The first line in the file is (format of every record in log file):

```
127.0.0.1 - - [25/Aug/2012:11:47:22 -0700] "GET / HTTP/1.1" 302 - "-" "Mozilla/5.0
(Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko)
Chrome/21.0.1180.83 Safari/537.1"
```

Log Record Details

Field	Field Name	Field Type
		IP_Address

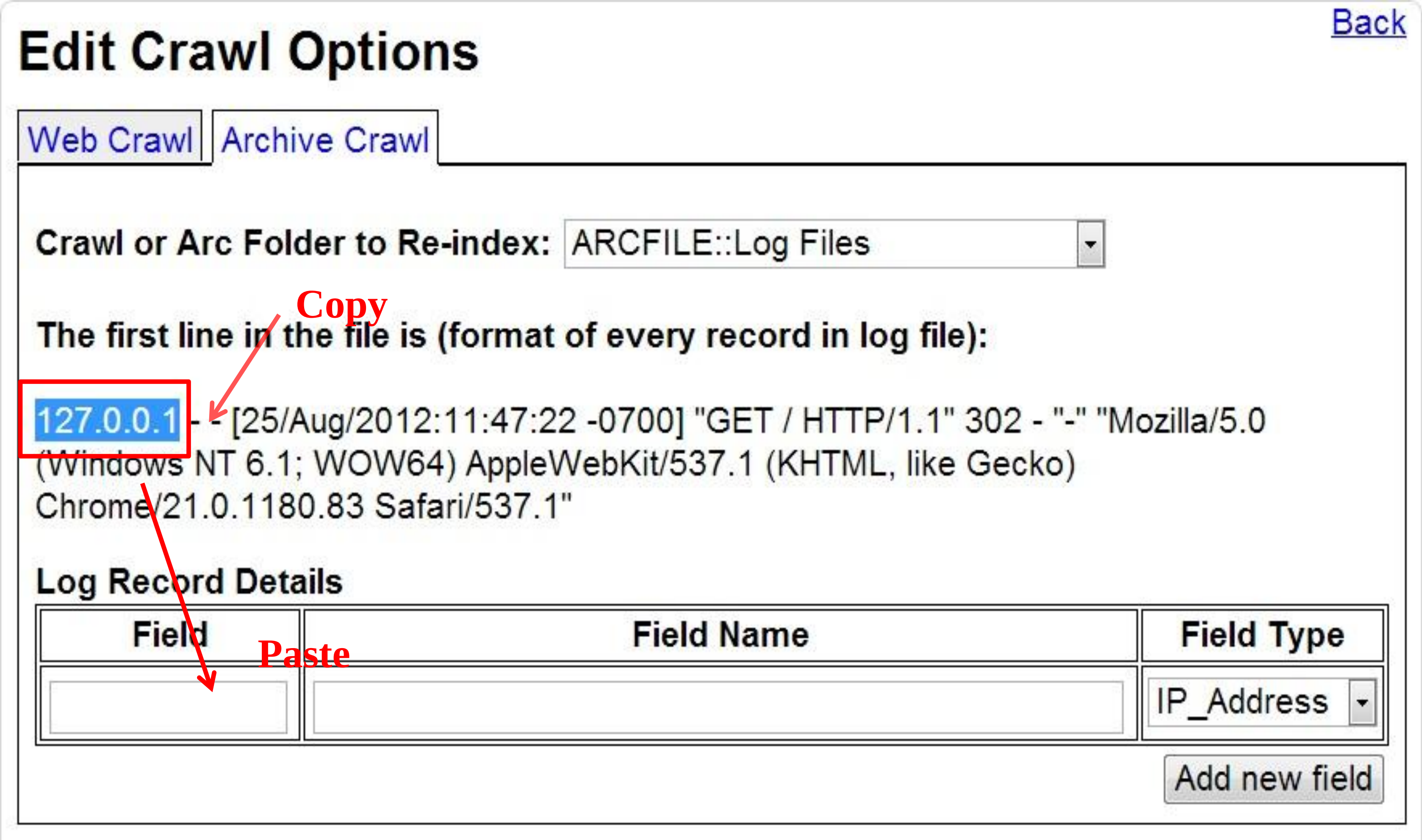
Add new field

Meta Words

Word	URL Pattern

Figure : User Interface for Log Files Crawling

Modifications to Yioop! (Cont…)



Edit Crawl Options

Back

Web Crawl | Archive Crawl

Crawl or Arc Folder to Re-index: ARCFILE::Log Files

The first line in the file is (format of every record in log file):

```
127.0.0.1 - - [25/Aug/2012:11:47:22 -0700] "GET / HTTP/1.1" 302 - "-" "Mozilla/5.0
(Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko)
Chrome/21.0.1180.83 Safari/537.1"
```

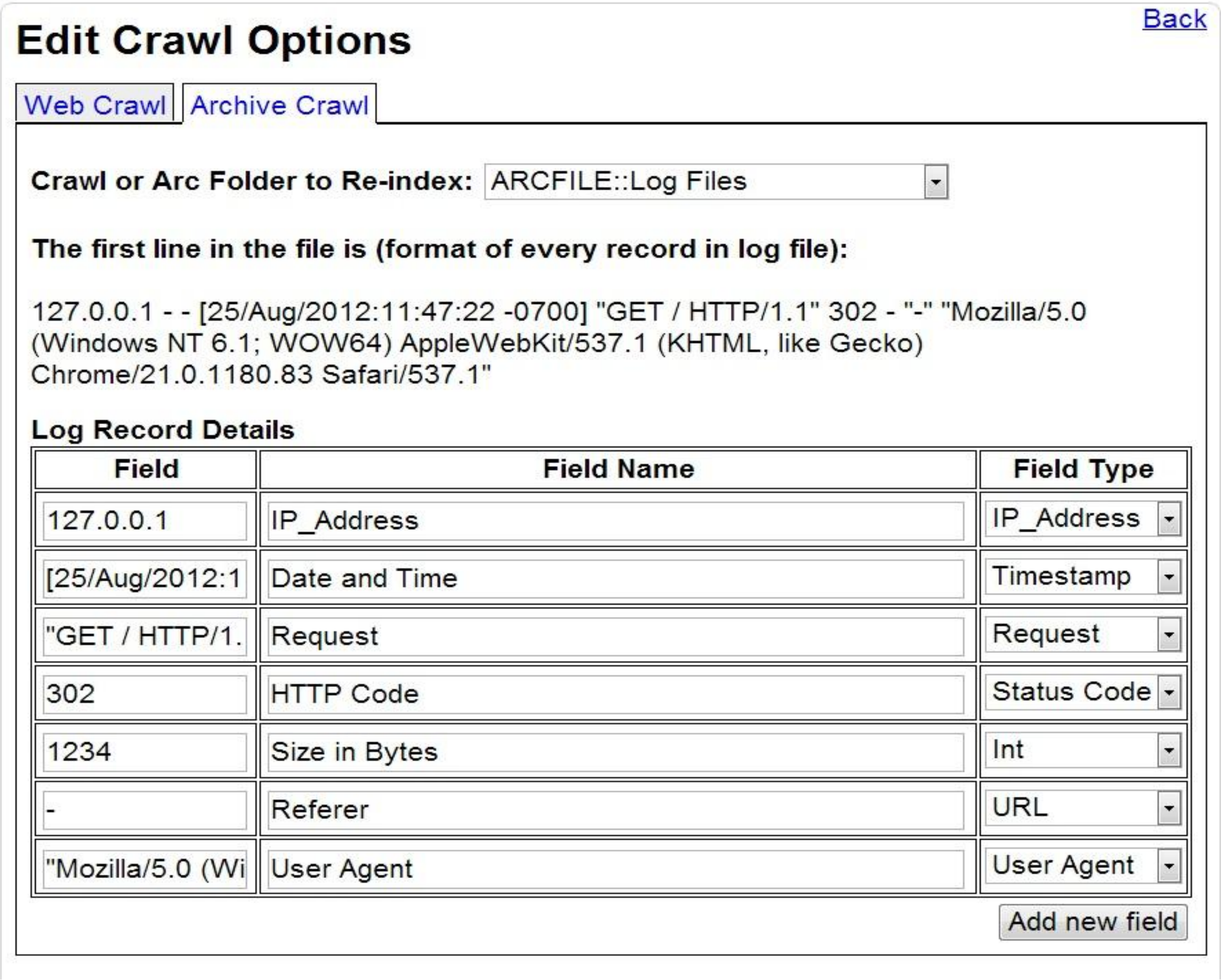
Log Record Details

Field	Field Name	Field Type
		IP_Address

Add new field

Figure : Entering details into the interface

Modifications to Yioop! (Cont…)



Edit Crawl Options

Crawl or Arc Folder to Re-index: ARCFILE::Log Files

The first line in the file is (format of every record in log file):

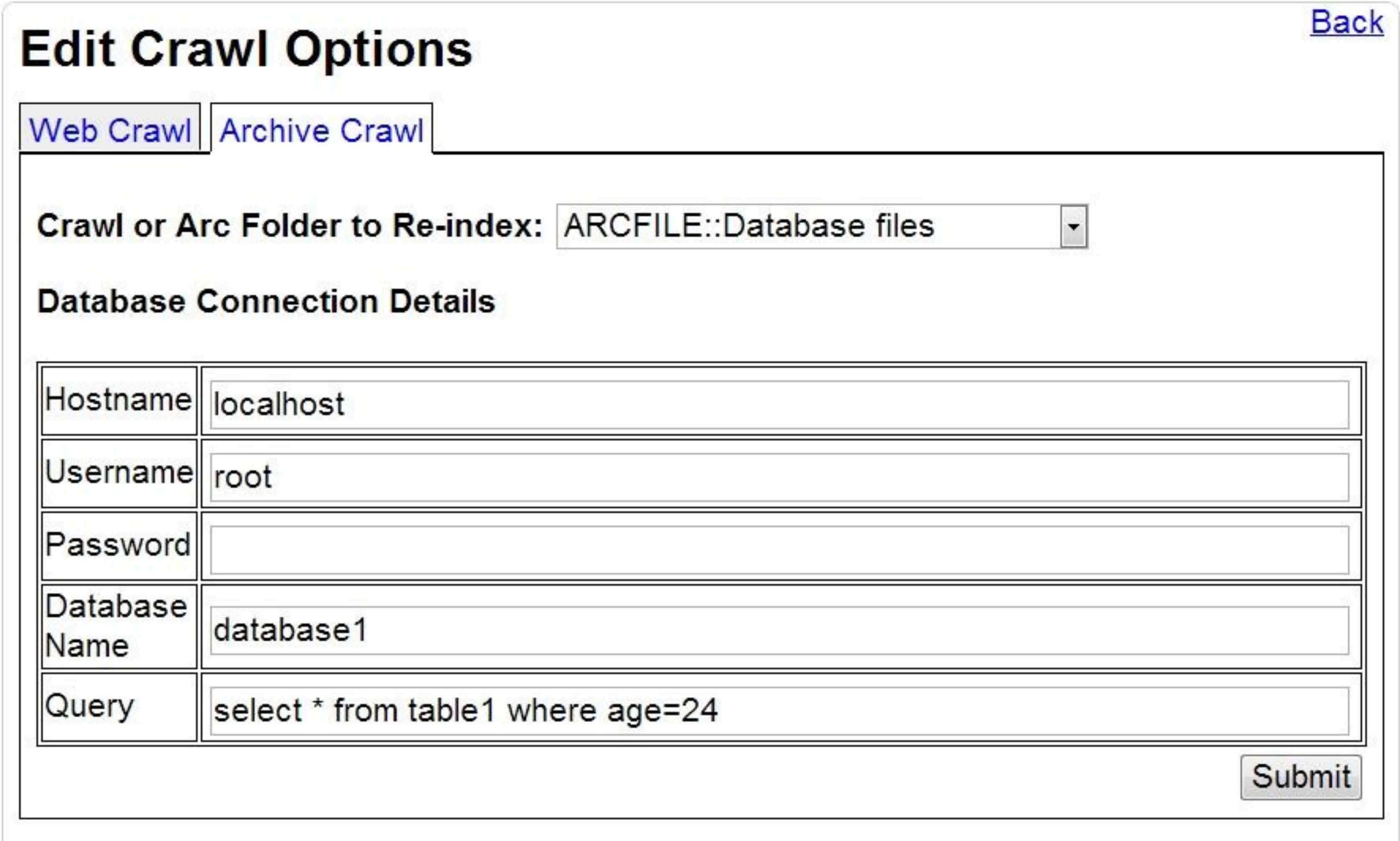
127.0.0.1 - - [25/Aug/2012:11:47:22 -0700] "GET / HTTP/1.1" 302 - "-" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.83 Safari/537.1"

Log Record Details

Field	Field Name	Field Type
127.0.0.1	IP_Address	IP_Address
[25/Aug/2012:1	Date and Time	Timestamp
"GET / HTTP/1.	Request	Request
302	HTTP Code	Status Code
1234	Size in Bytes	Int
-	Referer	URL
"Mozilla/5.0 (Wi	User Agent	User Agent

Figure : User Interface after inputting the details

Modifications to Yioop! (Cont…)



Edit Crawl Options

Crawl or Arc Folder to Re-index: ARCFILE::Database files

Database Connection Details

Hostname	localhost
Username	root
Password	
Database Name	database1
Query	select * from table1 where age=24

Figure : User Interface for crawling databases

Modifications to Yioop! (Cont…)

➢ In order to crawl and index the log records and databases in Yioop, two archive bundle iterator was added into the yioop/lib/archive_bundle_iterators folder.

➢The files that are added into the folder are namely,

- Log_archive_bundle_iterator.php
- Database_archive_bundle_iterator.php

➢ Both the archive bundle iterators contain few major functions:

➢**__construct**

- This is the constructor responsible for creating an instance of an archive iterator with timestamps and also for calling the appropriate functions.

Modifications to Yioop! (Cont…)

➢nextPages

- The nextPages function would call the nextPage function for every record to be crawled and converts it into a web page.

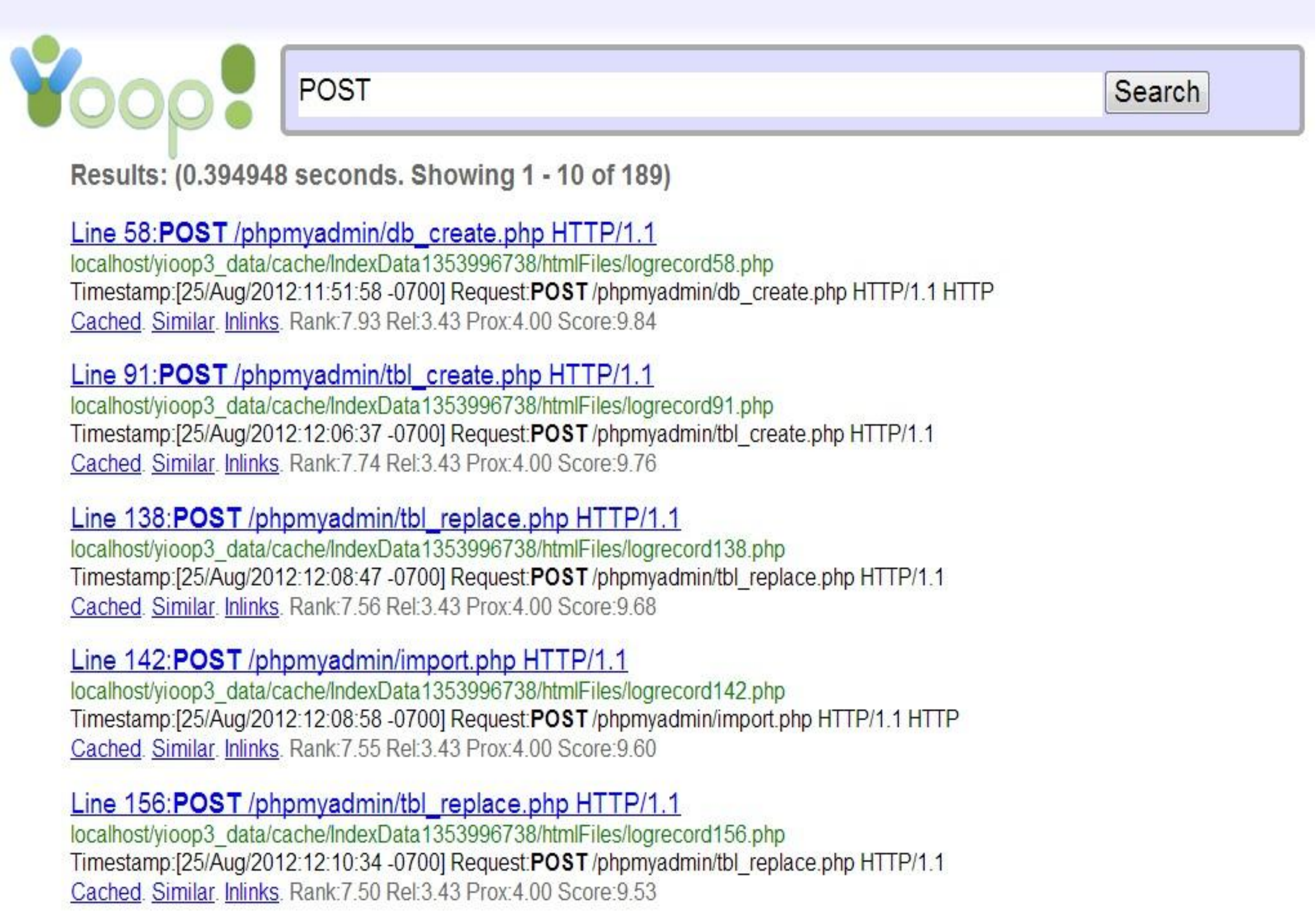
➢nextPage

- The nextPage function is responsible for creating the HTML pages of each log record.
- One can also add new text nodes to the web page in order to mention anything interesting about the log record.
- This one becomes an important function in the archive bundle iterators.

Modifications to Yioop! (Cont…)

- **parseLogRecord (Log files)**
  - This function is called for every log record and it parses the record according to the regular expressions and returns the results.
- **createRecords (Databases)**
  - This would create the database records after executing the query specified by the user.
- Apart from the functions explained here, there are many other functions performing different functions.

Modifications to Yioop! (Cont...)



Results: (0.394948 seconds. Showing 1 - 10 of 189)

Line 58:**POST** /phpmyadmin/db_create.php HTTP/1.1
localhost/yioop3_data/cache/IndexData1353996738/htmlFiles/logrecord58.php
Timestamp:[25/Aug/2012:11:51:58 -0700] Request:**POST** /phpmyadmin/db_create.php HTTP/1.1 HTTP
Cached. Similar. Inlinks. Rank:7.93 Rel:3.43 Prox:4.00 Score:9.84

Line 91:**POST** /phpmyadmin/tbl_create.php HTTP/1.1
localhost/yioop3_data/cache/IndexData1353996738/htmlFiles/logrecord91.php
Timestamp:[25/Aug/2012:12:06:37 -0700] Request:**POST** /phpmyadmin/tbl_create.php HTTP/1.1
Cached. Similar. Inlinks. Rank:7.74 Rel:3.43 Prox:4.00 Score:9.76

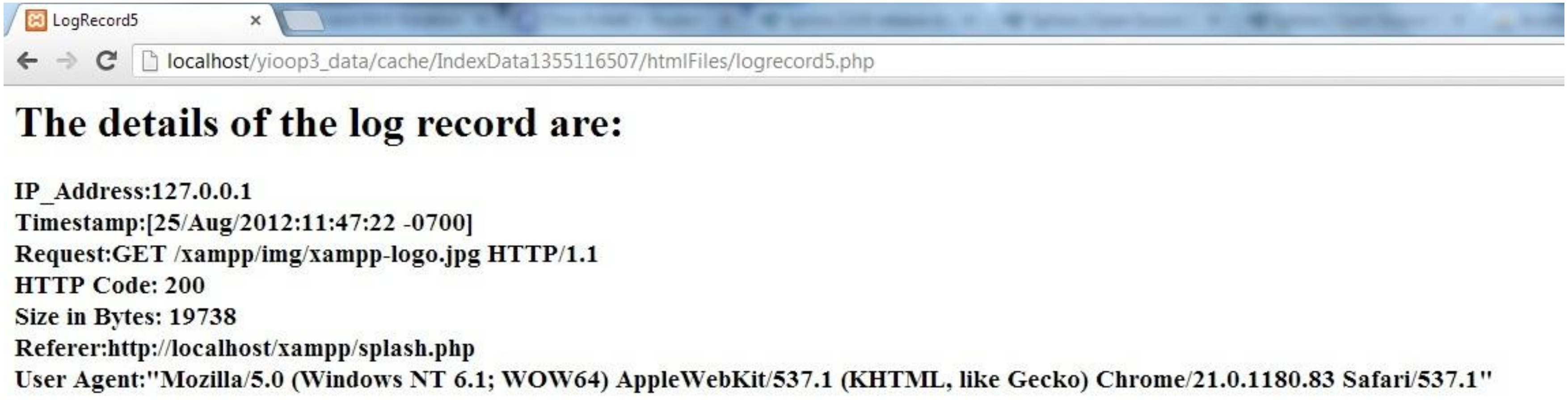
Line 138:**POST** /phpmyadmin/tbl_replace.php HTTP/1.1
localhost/yioop3_data/cache/IndexData1353996738/htmlFiles/logrecord138.php
Timestamp:[25/Aug/2012:12:08:47 -0700] Request:**POST** /phpmyadmin/tbl_replace.php HTTP/1.1
Cached. Similar. Inlinks. Rank:7.56 Rel:3.43 Prox:4.00 Score:9.68

Line 142:**POST** /phpmyadmin/import.php HTTP/1.1
localhost/yioop3_data/cache/IndexData1353996738/htmlFiles/logrecord142.php
Timestamp:[25/Aug/2012:12:08:58 -0700] Request:**POST** /phpmyadmin/import.php HTTP/1.1 HTTP
Cached. Similar. Inlinks. Rank:7.55 Rel:3.43 Prox:4.00 Score:9.60

Line 156:**POST** /phpmyadmin/tbl_replace.php HTTP/1.1
localhost/yioop3_data/cache/IndexData1353996738/htmlFiles/logrecord156.php
Timestamp:[25/Aug/2012:12:10:34 -0700] Request:**POST** /phpmyadmin/tbl_replace.php HTTP/1.1
Cached. Similar. Inlinks. Rank:7.50 Rel:3.43 Prox:4.00 Score:9.53

Figure : Search results obtained after crawling log records

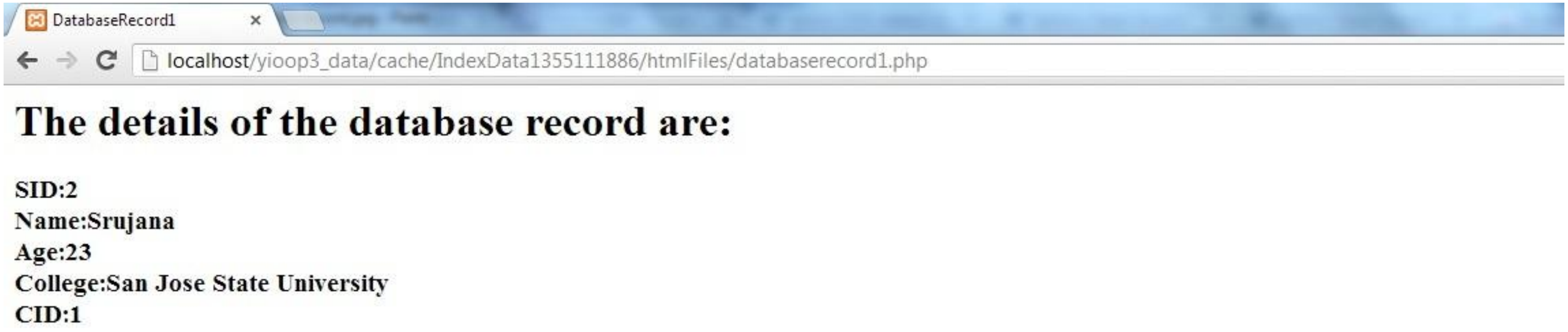
Modifications to Yioop! (Cont…)


LogRecord5
localhost/yioop3_data/cache/IndexData1355116507/htmlFiles/logrecord5.php

The details of the log record are:

IP_Address:127.0.0.1
Timestamp:[25/Aug/2012:11:47:22 -0700]
Request:GET /xampp/img/xampp-logo.jpg HTTP/1.1
HTTP Code: 200
Size in Bytes: 19738
Referer:http://localhost/xampp/splash.php
User Agent:"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.83 Safari/537.1"

Figure : An open log record


DatabaseRecord1
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord1.php

The details of the database record are:

SID:2
Name:Srujana
Age:23
College:San Jose State University
CID:1

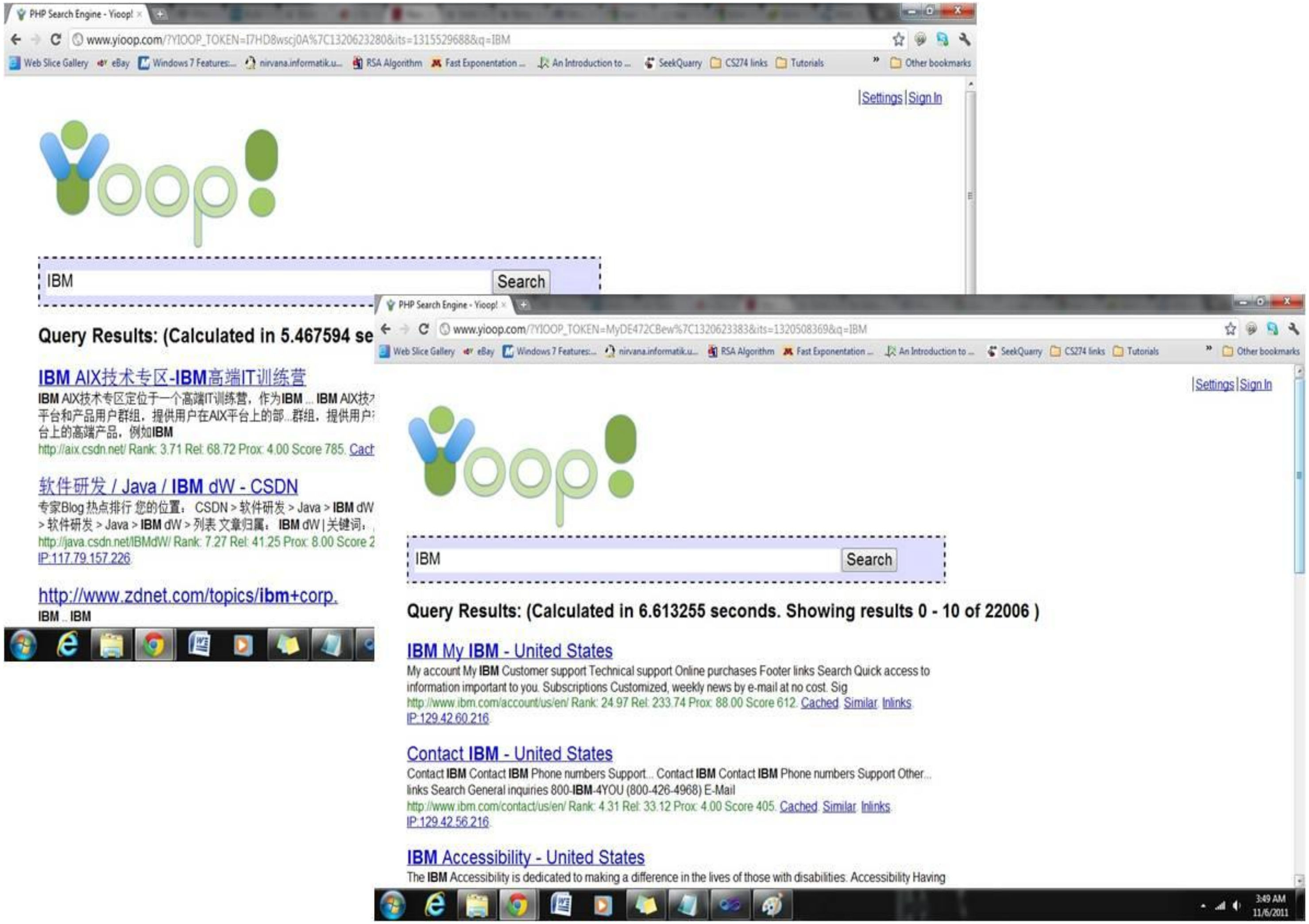
Figure : An open database record

Modifications to Yioop! (Cont…)

In order to make Yioop work correctly with the short links, major part of the code was added in the following files of Yioop.

- **fetch_url.php**
  - The extraction of the "Location" information from the short URL is done in this file.
  - This information was required to point the results to the original URL and not the short URL.
- **fetcher.php**
  - This step changes the short URL to the original link using the location information extracted from fetch_url.php.

Modifications to Yioop! (Cont…)



IBM

Query Results: (Calculated in 5.467594 se

IBM AIX技术专区-IBM高端IT训练营

IBM AIX技术专区定位于一个高端IT训练营，作为**IBM** ... **IBM** AIX技 平台和产品用户群组，提供用户在AIX平台上的部...群组，提供用户 台上的高端产品，例如**IBM**

http://aix.csdn.net/ Rank: 3.71 Rel: 68.72 Prox: 4.00 Score 785. Cach

软件研发 / Java / IBM dW - CSDN

专家Blog 热点排行 您的位置： CSDN > 软件研发 > Java > **IBM** dW > 软件研发 > Java > **IBM** dW > 列表 文章归属： **IBM** dW | 关键词：

http://java.csdn.net/IBMdW/ Rank: 7.27 Rel: 41.25 Prox: 8.00 Score 2

IP:117.79.157.226

http://www.zdnet.com/topics/ibm+corp.

IBM .. **IBM**

IBM

Query Results: (Calculated in 6.613255 seconds. Showing results 0 - 10 of 22006)

IBM My IBM - United States

My account My **IBM** Customer support Technical support Online purchases Footer links Search Quick access to information important to you. Subscriptions Customized, weekly news by e-mail at no cost. Sig

http://www.ibm.com/account/us/en/ Rank: 24.97 Rel: 233.74 Prox: 88.00 Score 612. Cached. Similar. Inlinks.

IP:129.42.60.216

Contact IBM - United States

Contact **IBM** Contact **IBM** Phone numbers Support... Contact **IBM** Contact **IBM** Phone numbers Support Other... links Search General inquiries 800-**IBM**-4YOU (800-426-4968) E-Mail

http://www.ibm.com/contact/us/en/ Rank: 4.31 Rel: 33.12 Prox: 4.00 Score 405. Cached. Similar. Inlinks.

IP:129.42.56.216

IBM Accessibility - United States

The **IBM** Accessibility is dedicated to making a difference in the lives of those with disabilities. Accessibility Having

Figure : Results before and after the code was added

Tests and Results

➢Testing has been performed in order to test the following:

- Usability of the user interface developed in Yioop
- Efficiency of the archive bundle iterators
- Various Database queries
- Improvement in results after code is added for short links.

➢The testing methods and results are described in the following slides.

Usability testing

➢Usability testing is a technique used in the field of user interface design or user centered design to evaluate a product by testing it on users [5].

➢Usability testing is done to measure to what extent the user is satisfied in the following four subject areas :

- **Efficiency** - The time taken or the number of steps taken for a user to complete a particular task.
- **Accuracy** - The number of mistakes made by a user while performing a particular task.
- **Recall** - This measure refers to the amount of work that a user could recall after a period of time concerned with a particular task.
- **Emotional Response** - This refers to the amount of satisfaction and the kind of emotion felt by a person while performing a particular task.

Usability testing (Cont...)

➢In order to perform testing, there were seven tasks that the user was asked to perform.

- **Task1 :** Download and Install Yioop (Includes downloading and installing Xampp)
- **Task2 :** Configure Yioop
- **Task3 :** Start a new crawl in Yioop
- **Task4 :** Set up log files folder in Yioop for archive crawling
- **Task5 :** Set up a folder for database in Yioop
- **Task6 :** Input details into log files interface, save them and start a crawl with them
- **Task7 :** Input details into database interface, save them and start a crawl with them

Efficiency

➢The time taken for each user to complete each task is recorded.

➢Most users took the same amount of time to complete each of the tasks.

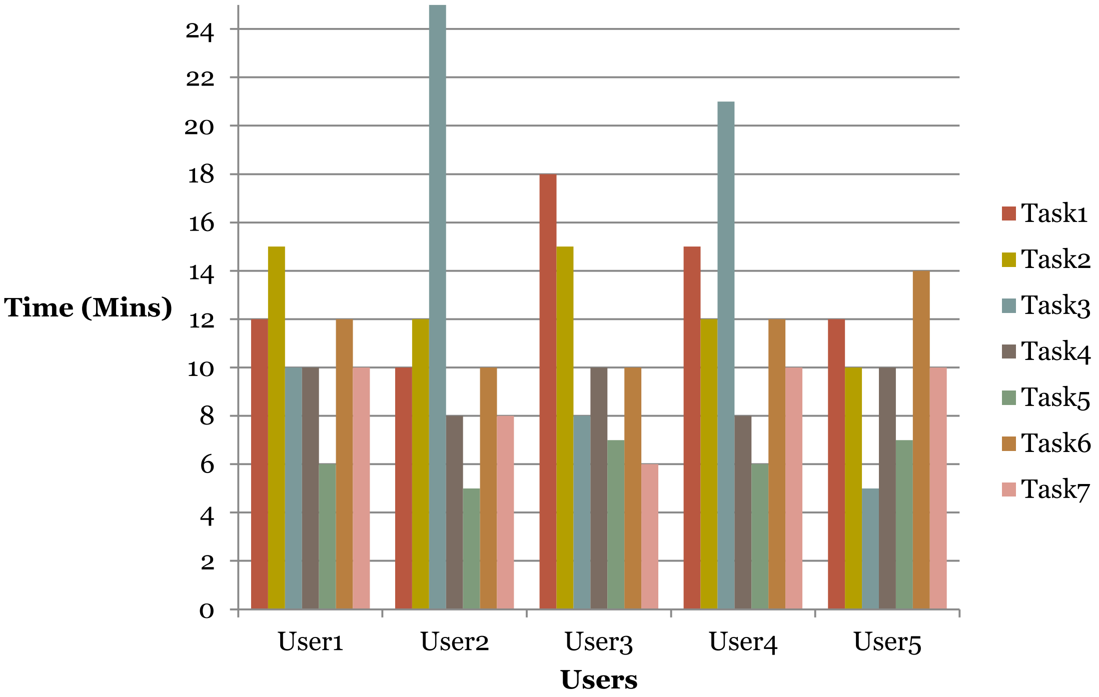


Figure : Time taken for each user to complete a task

Efficiency (Cont...)

- The average amount of time taken for each of the tasks by the users is calculated.

- It gives a good comparison between the amount of time taken by normal Yioop processes and the new features implemented.

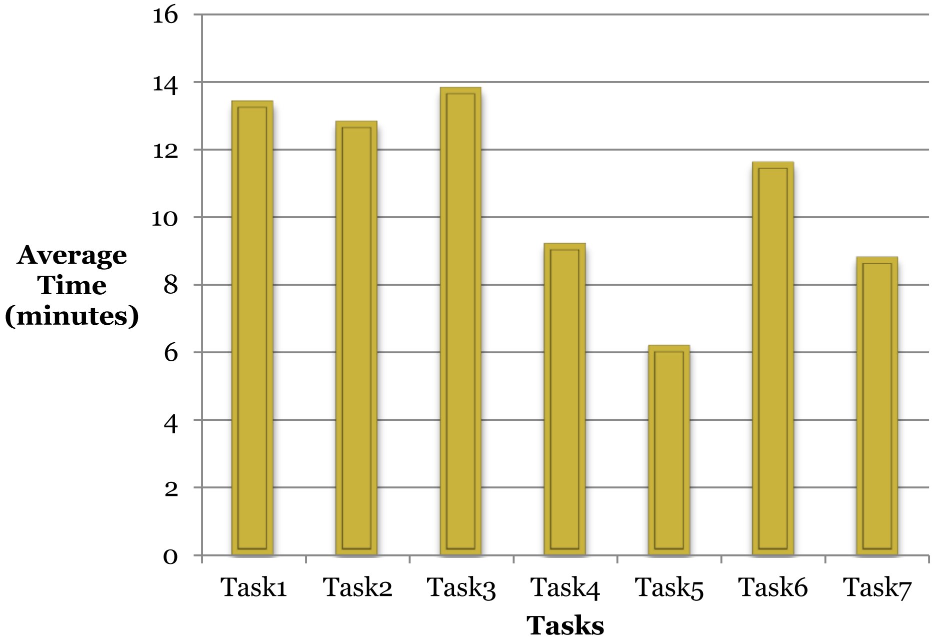
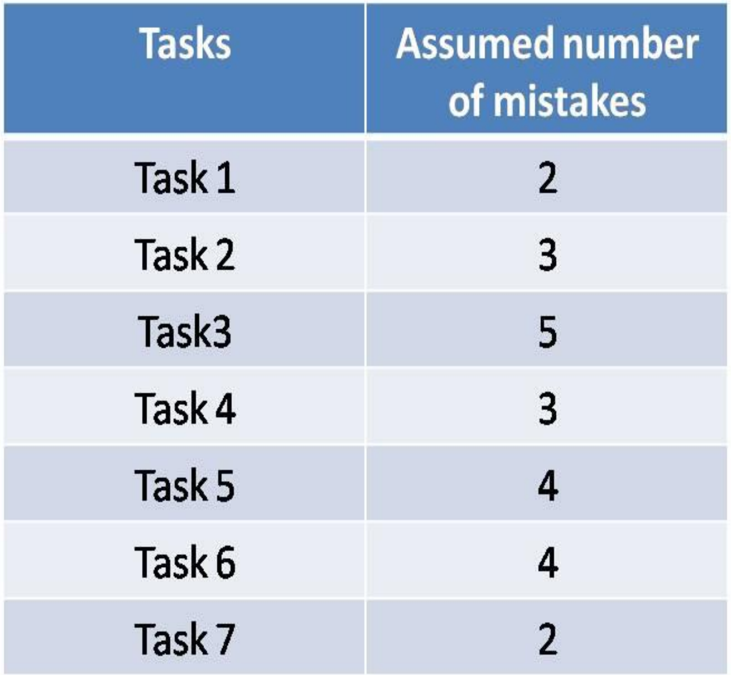


Figure : Average time taken by users for each of the tasks

Accuracy

- The average number of mistakes made by the users for a particular task are calculated.

- A certain number of mistakes are assumed to take place for every task.



Tasks	Assumed number of mistakes
Task 1	2
Task 2	3
Task3	5
Task 4	3
Task 5	4
Task 6	4
Task 7	2

Table : Assumed number of mistakes for each task

Accuracy (Cont…)

- The average number of mistakes done are compared over the assumed number of mistakes and the accuracy percentage is calculated.

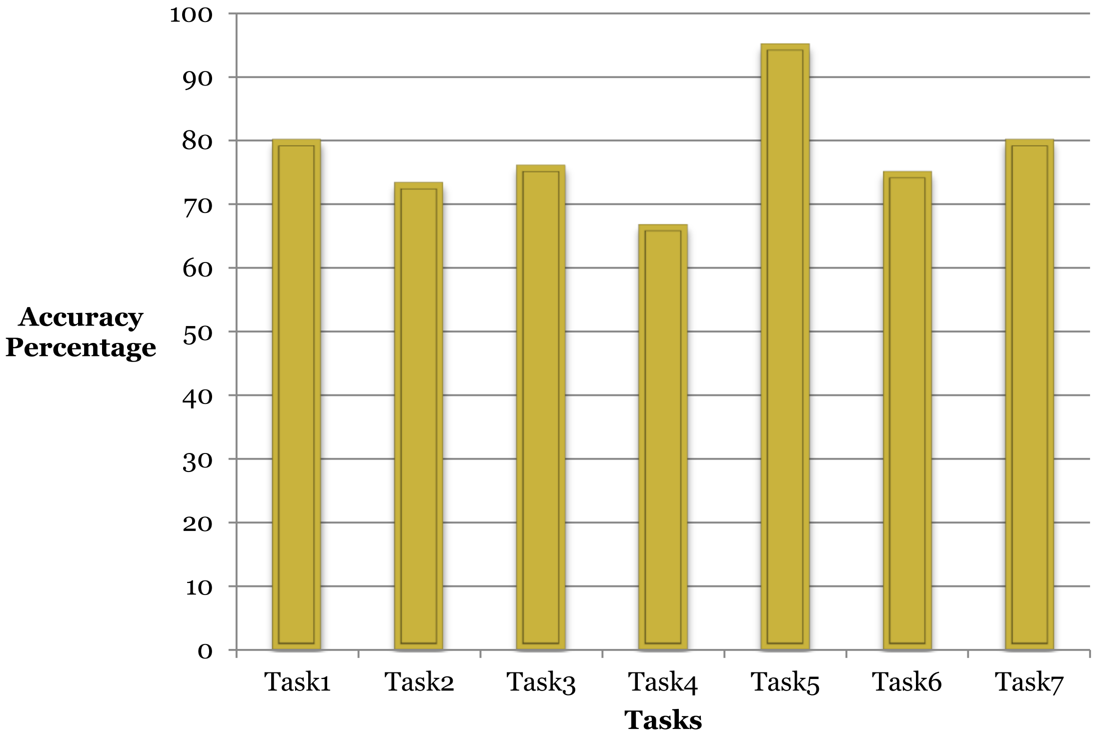


Figure : Accuracy percentage of tasks performed by the user

Recall

- It is calculated by comparing the amount of time taken by the user to do the tasks for the first time and the second time.

- A decrease in the amount of time shows improvement in the tasks performed.

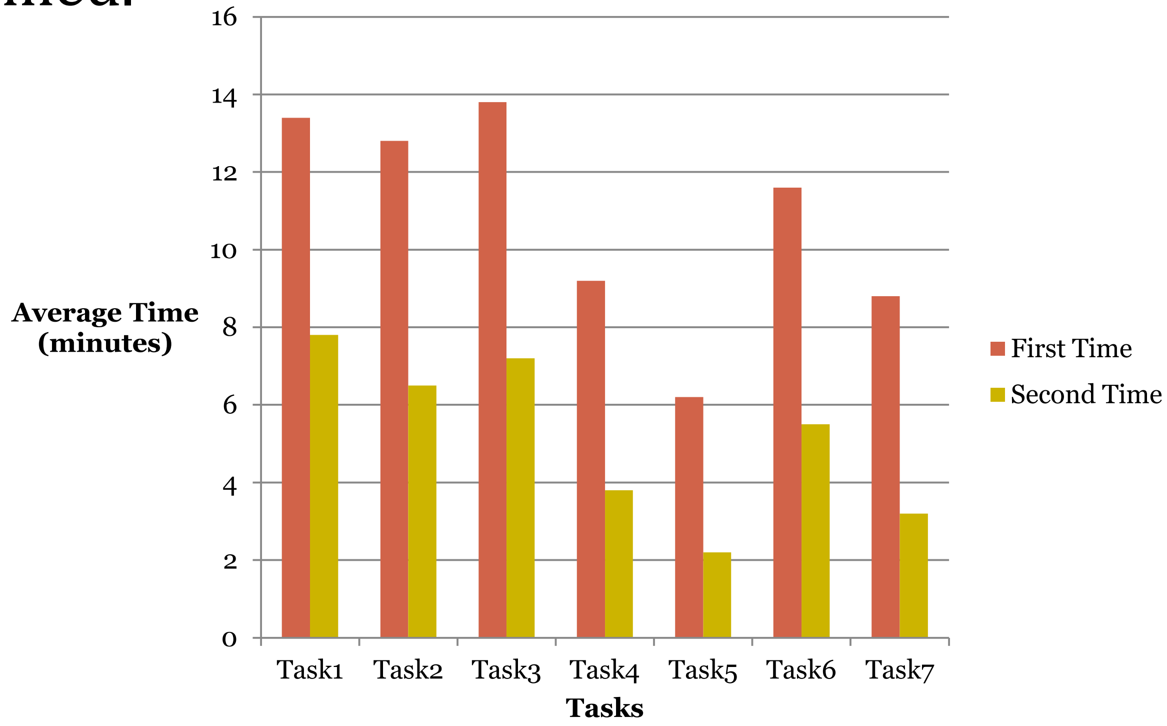


Figure : Time taken by users for the first time and second time

Emotional Response

- Each user was asked to give a rating based on their satisfaction levels.

- The rating was on a scale of 1 to 10, where 1 is completely dissatisfied and 10 is completely satisfied.

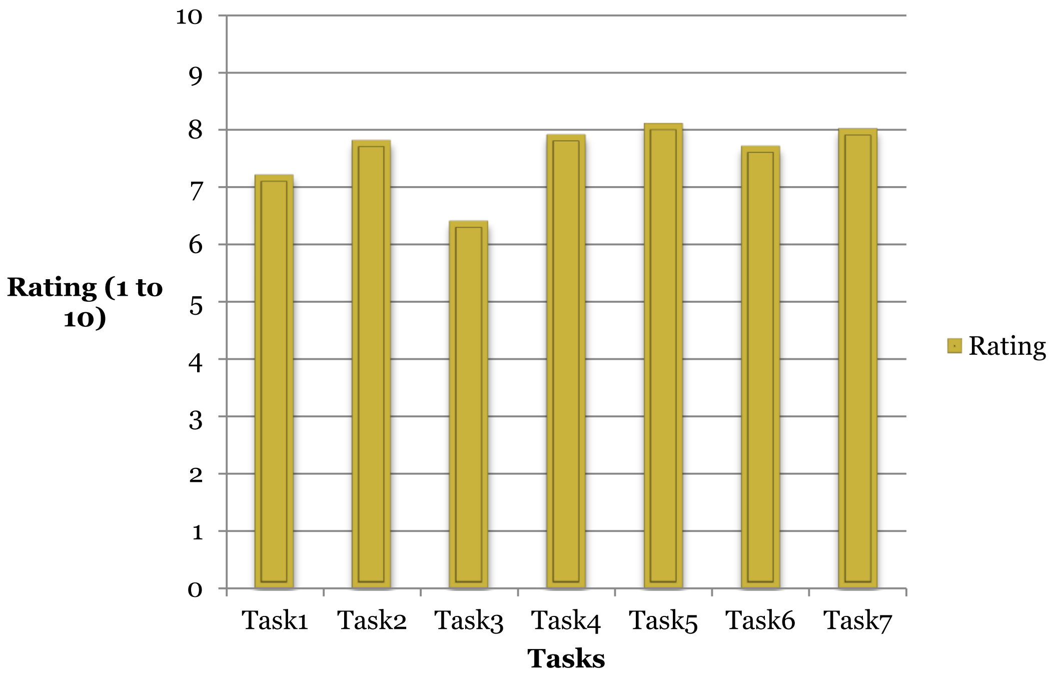


Figure : Average rating of the satisfaction of user

Testing Efficiency

➢The log archive bundle iterator was tested by varying the number of records ranging from 10000 record to 100,000 records.

➢These crawls were performed using a single machine with a single server. The times taken for each of these crawls is recorded.

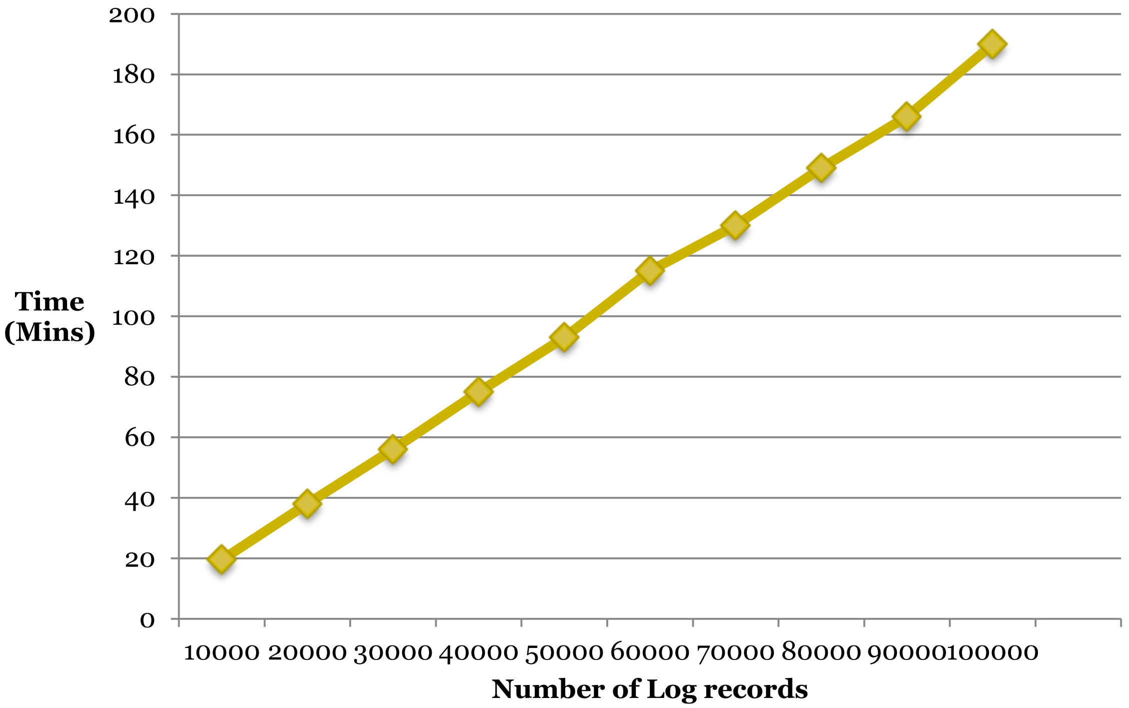


Figure : Exponential growth in time with the increase in the number of log records

Testing Efficiency (Cont…)

- The database archive iterator was tested by providing different queries in the database interface.

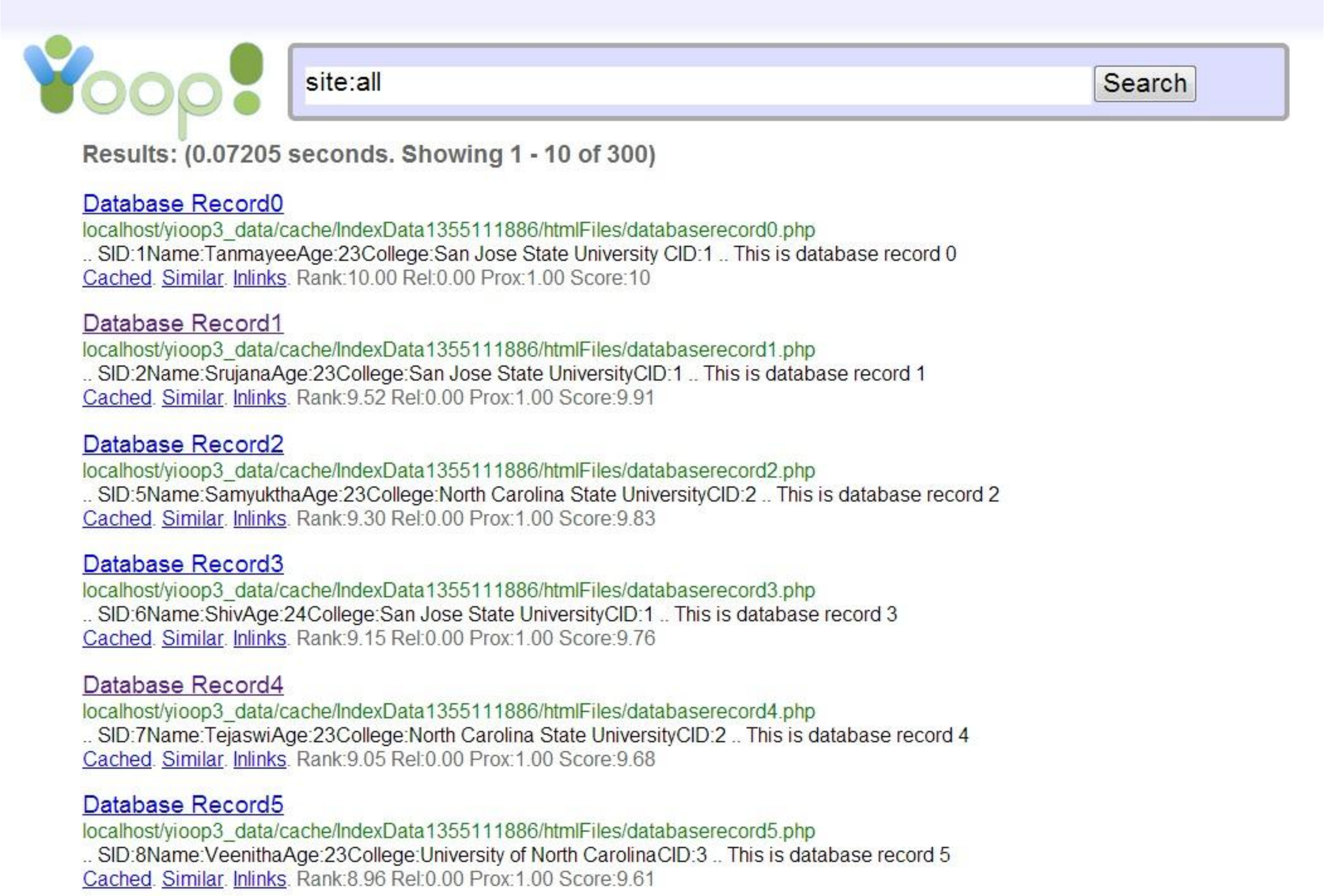
- It was checked if the records are limited and extracted based on the queries inputted.

- The database table used for this purpose is populated with 20000 records.

- The first query inputted was to query the entire table.

- The second query inputted was to crawl only those records in the table whose age field is 24.

Testing Efficiency (Cont...)



Results: (0.07205 seconds. Showing 1 - 10 of 300)

Database Record0
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord0.php
.. SID:1Name:TanmayeeAge:23College:San Jose State University CID:1 .. This is database record 0
Cached. Similar. Inlinks. Rank:10.00 Rel:0.00 Prox:1.00 Score:10

Database Record1
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord1.php
.. SID:2Name:SrujanaAge:23College:San Jose State UniversityCID:1 .. This is database record 1
Cached. Similar. Inlinks. Rank:9.52 Rel:0.00 Prox:1.00 Score:9.91

Database Record2
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord2.php
.. SID:5Name:SamyukthaAge:23College:North Carolina State UniversityCID:2 .. This is database record 2
Cached. Similar. Inlinks. Rank:9.30 Rel:0.00 Prox:1.00 Score:9.83

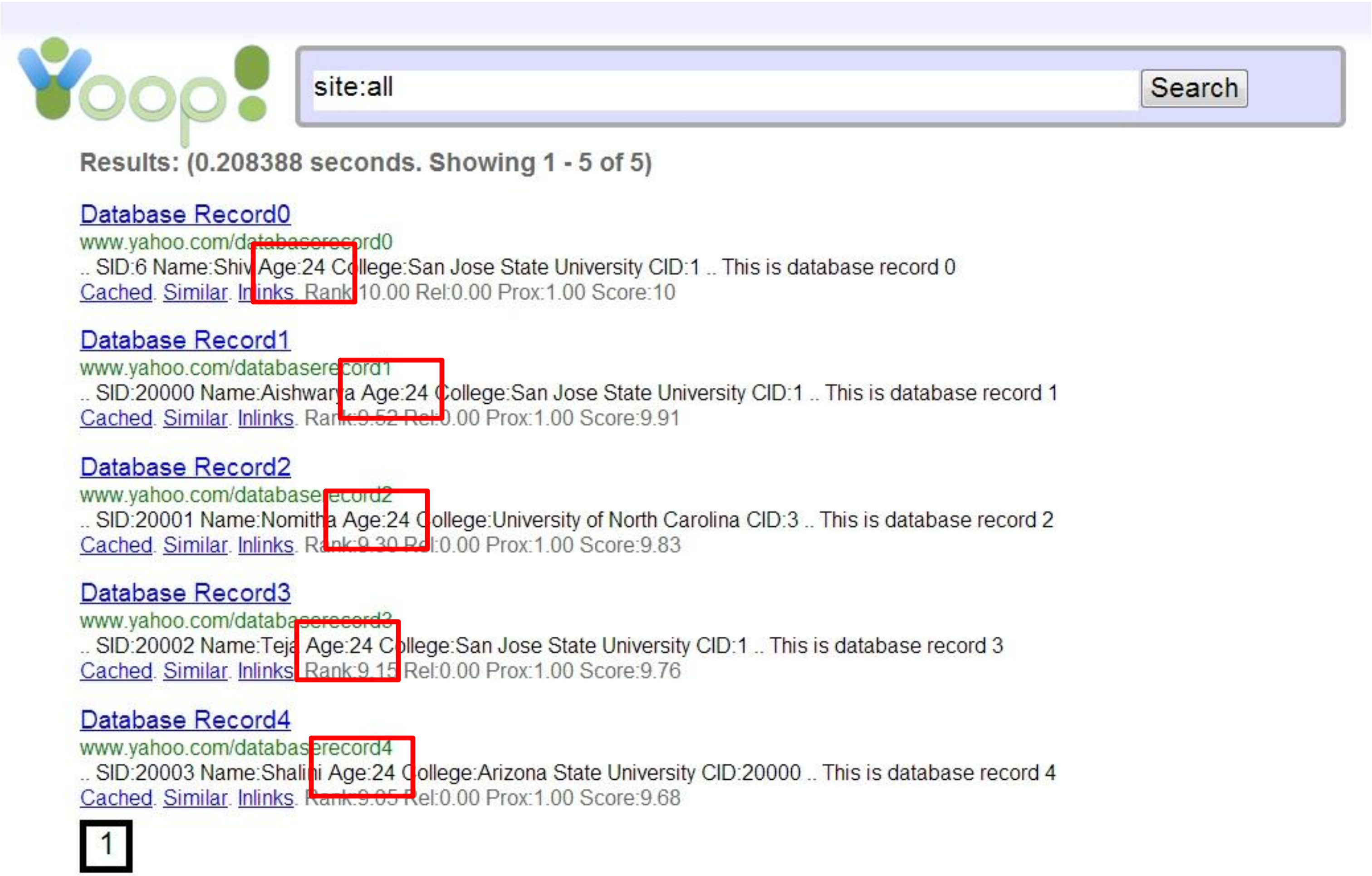
Database Record3
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord3.php
.. SID:6Name:ShivAge:24College:San Jose State UniversityCID:1 .. This is database record 3
Cached. Similar. Inlinks. Rank:9.15 Rel:0.00 Prox:1.00 Score:9.76

Database Record4
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord4.php
.. SID:7Name:TejaswiAge:23College:North Carolina State UniversityCID:2 .. This is database record 4
Cached. Similar. Inlinks. Rank:9.05 Rel:0.00 Prox:1.00 Score:9.68

Database Record5
localhost/yioop3_data/cache/IndexData1355111886/htmlFiles/databaserecord5.php
.. SID:8Name:VeenithaAge:23College:University of North CarolinaCID:3 .. This is database record 5
Cached. Similar. Inlinks. Rank:8.96 Rel:0.00 Prox:1.00 Score:9.61

Figure : Crawl results for first query

Testing Efficiency (Cont...)



Results: (0.208388 seconds. Showing 1 - 5 of 5)

Database Record0
www.yahoo.com/databaserecord0
.. SID:6 Name:Shiva Age:24 College:San Jose State University CID:1 .. This is database record 0
Cached. Similar. Inlinks. Rank:10.00 Rel:0.00 Prox:1.00 Score:10

Database Record1
www.yahoo.com/databaserecord1
.. SID:20000 Name:Aishwarya Age:24 College:San Jose State University CID:1 .. This is database record 1
Cached. Similar. Inlinks. Rank:9.52 Rel:0.00 Prox:1.00 Score:9.91

Database Record2
www.yahoo.com/databaserecord2
.. SID:20001 Name:Nomitha Age:24 College:University of North Carolina CID:3 .. This is database record 2
Cached. Similar. Inlinks. Rank:9.30 Rel:0.00 Prox:1.00 Score:9.83

Database Record3
www.yahoo.com/databaserecord3
.. SID:20002 Name:Teja Age:24 College:San Jose State University CID:1 .. This is database record 3
Cached. Similar. Inlinks. Rank:9.15 Rel:0.00 Prox:1.00 Score:9.76

Database Record4
www.yahoo.com/databaserecord4
.. SID:20003 Name:Shalini Age:24 College:Arizona State University CID:20000 .. This is database record 4
Cached. Similar. Inlinks. Rank:9.05 Rel:0.00 Prox:1.00 Score:9.68

Figure : Crawl results for second query

URL Shortening Testing

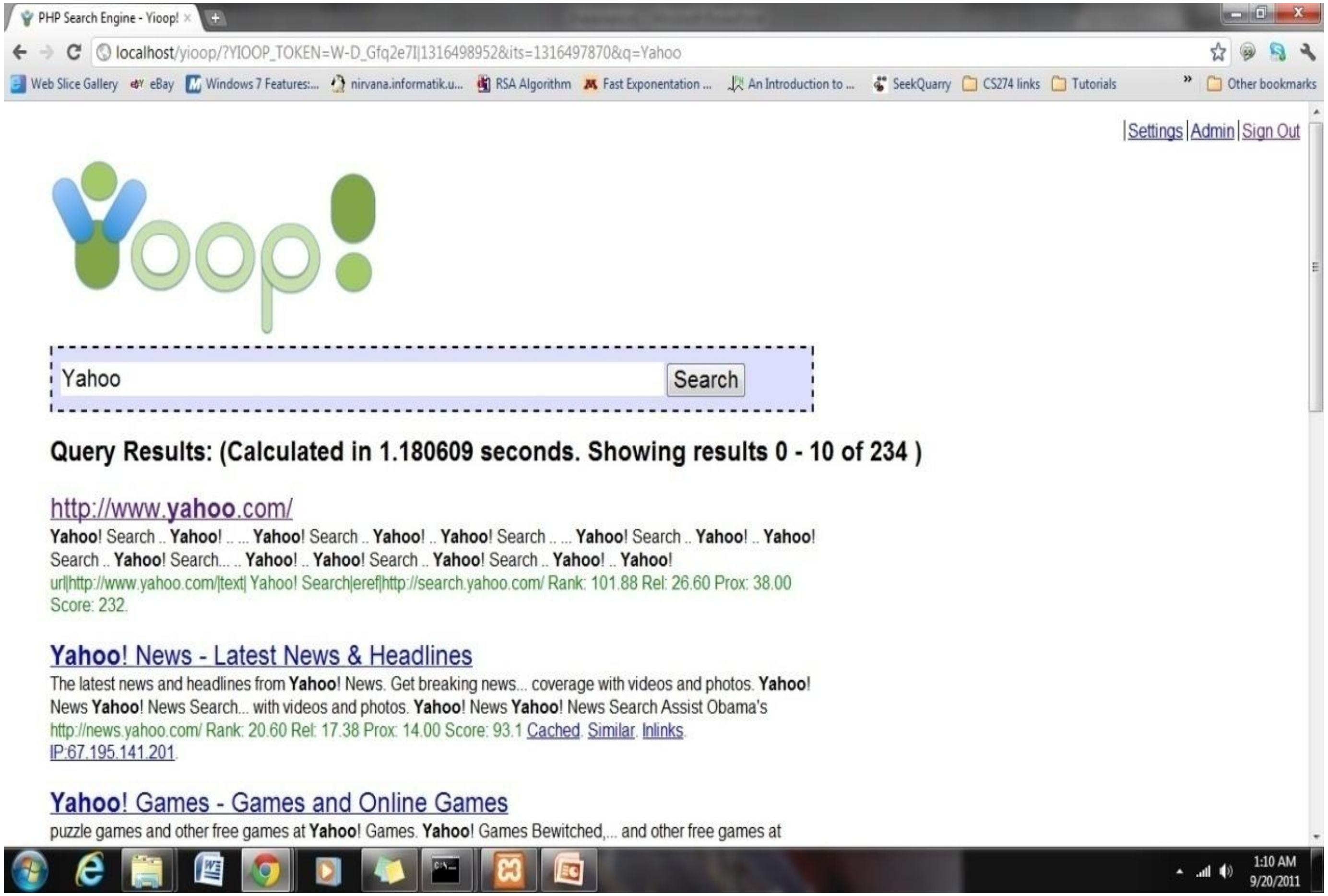
- In order to test if Yioop works correctly with short links, it was tested to see if the results improved after the code is embedded into Yioop.

- Firstly, a crawl is performed on a bitly link. This crawl is performed before adding the code and also after adding the code.

- The results were compared and it showed that Yioop improved on the results displayed.

- Also, the short links were avoided and redirected to the original links and these links were crawled and indexed.

URL Shortening Testing (Cont…)



Query Results: (Calculated in 1.180609 seconds. Showing results 0 - 10 of 234)

http://www.**yahoo**.com/

Yahoo! Search .. **Yahoo**! .. ... **Yahoo**! Search .. **Yahoo**! .. **Yahoo**! Search .. ... **Yahoo**! Search .. **Yahoo**! .. **Yahoo**! Search .. **Yahoo**! Search... .. **Yahoo**! .. **Yahoo**! Search .. **Yahoo**! Search .. **Yahoo**! .. **Yahoo**!

url|http://www.yahoo.com/|text| Yahoo! Search|eref|http://search.yahoo.com/ Rank: 101.88 Rel: 26.60 Prox: 38.00 Score: 232.

Yahoo! News - Latest News & Headlines

The latest news and headlines from **Yahoo**! News. Get breaking news... coverage with videos and photos. **Yahoo**! News **Yahoo**! News Search... with videos and photos. **Yahoo**! News **Yahoo**! News Search Assist Obama's

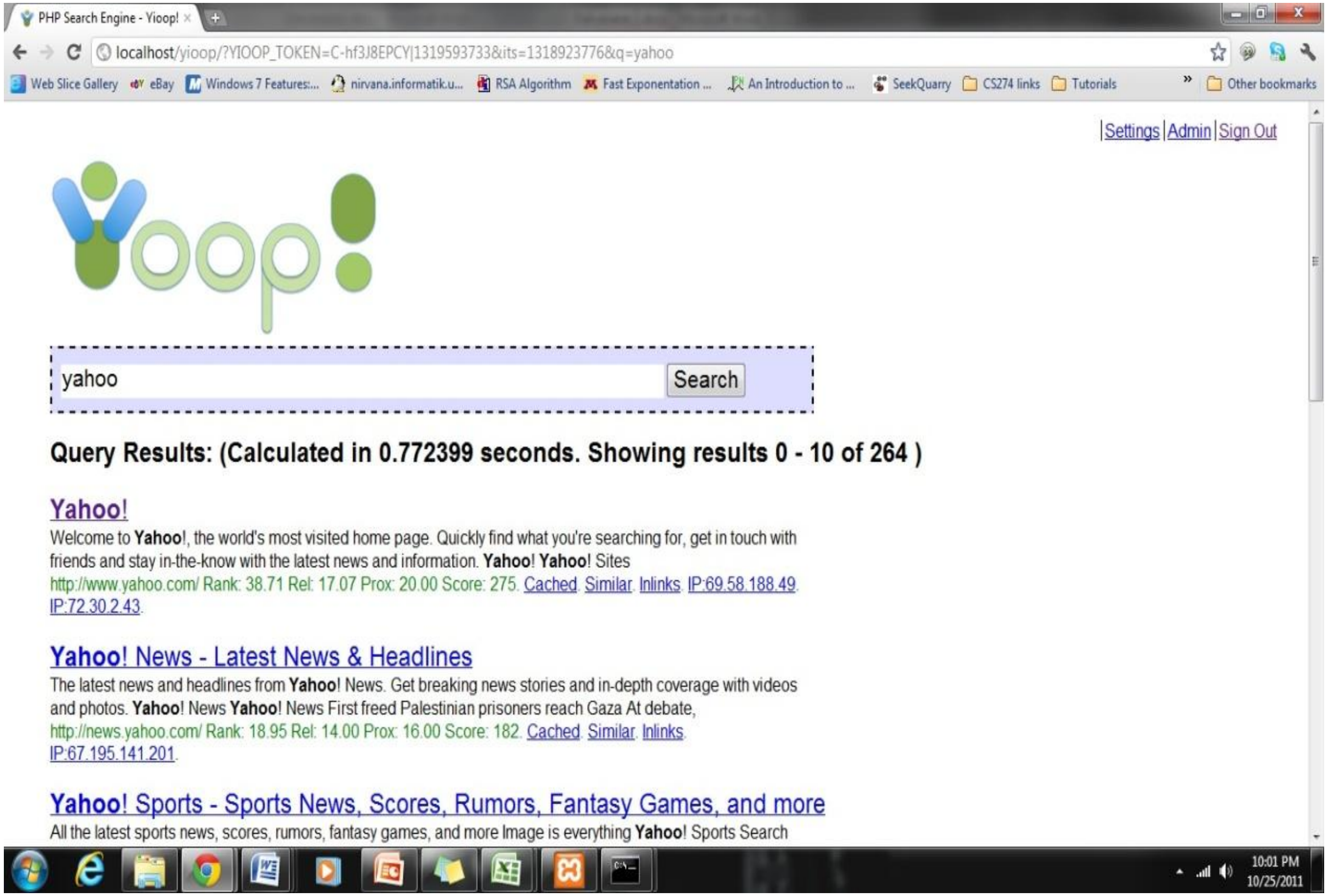
http://news.yahoo.com/ Rank: 20.60 Rel: 17.38 Prox: 14.00 Score: 93.1 Cached. Similar. Inlinks. IP:67.195.141.201.

Yahoo! Games - Games and Online Games

puzzle games and other free games at **Yahoo**! Games. **Yahoo**! Games Bewitched,... and other free games at

Figure : Results before addition of the code

URL Shortening Testing (Cont…)



localhost/yioop/?YIOOP_TOKEN=C-hf3J8EPCY|1319593733&its=1318923776&q=yahoo

yahoo

Query Results: (Calculated in 0.772399 seconds. Showing results 0 - 10 of 264)

Yahoo!

Welcome to **Yahoo**!, the world's most visited home page. Quickly find what you're searching for, get in touch with friends and stay in-the-know with the latest news and information. **Yahoo**! **Yahoo**! Sites

http://www.yahoo.com/ Rank: 38.71 Rel: 17.07 Prox: 20.00 Score: 275. Cached. Similar. Inlinks. IP:69.58.188.49. IP:72.30.2.43.

Yahoo! News - Latest News & Headlines

The latest news and headlines from **Yahoo**! News. Get breaking news stories and in-depth coverage with videos and photos. **Yahoo**! News **Yahoo**! News First freed Palestinian prisoners reach Gaza At debate,

http://news.yahoo.com/ Rank: 18.95 Rel: 14.00 Prox: 16.00 Score: 182. Cached. Similar. Inlinks. IP:67.195.141.201.

Yahoo! Sports - Sports News, Scores, Rumors, Fantasy Games, and more

All the latest sports news, scores, rumors, fantasy games, and more Image is everything **Yahoo**! Sports Search

Figure : Results after the addition of the code

URL Shortening Testing

➢As seen in the figures,

- Before the addition of code, Yioop displayed the Yahoo URL as a plain text in the title of the page.

- After the addition of code, Yioop had redirected the bitly link correctly and the Yahoo link is displayed as the title and as the first result.

Conclusion

➢ Often, users might want to crawl and index different file formats.

➢User Interface created for user to perform customized crawling of log files and databases.

➢ Two archive bundle iterators written to crawl and index log files and databases.

➢ URL shortened services made to work appropriately in Yioop and found improved search results.

➢ Usability testing performed has shown user satisfaction.

References

(1) Pollett, C. (2012). Yioop! Documentation v 0.90. Retrieved from http://www.seekquarry.com/?c=main&p=documentation

(2) Peng, Wei., Tao, L., & Ma, S. (2005). Mining logs files for data-driven system management. *ACMSIGKDD Explorations Newsletter - Natural language processing and text mining, 7*(1).

(3) He, B., Patel, M., Zhang, Z., & Chang, K.C. (2007). Accessing the Deep Web. *ACM, 50*(5), 94-101.

(4) Deep Web. (2012, November 9). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Deep_Web

(5) Usability Testing. (2012, October 23). In *Wikipedia*. Retrieved from http://en.wikipedia.org/wiki/Deep_Web

References (Cont…)

(6) Bailey, B. (2006, March). Getting the complete picture with Usability Testing. Retrieved from http://www.usability.gov/articles/newsletter/pubs/030106news.html

(7) Lewandowski, D., & Mayr, P. (2006). Exploring the academic invisible web. *Library Hi Tech*, *24*(4), 529-539.

(8) Sphinx Documentation. (n.d.). In *Sphinx* website. Retrieved from http://sphinxsearch.com/docs/

(9) Server Log. (2012, July 18). In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Server_log

(10) Destailleur, L. (2008, December). AWStats. Retrieved from http://awstats.sourceforge.net/

(11) Log Files. (2012). In *Apache* website. Retrieved from http://httpd.apache.org/docs/1.3/logs.html

References (Cont...)

(12) Databases. (2012, November 26). In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Database

(13) Zillman, P. M. (2012, November 1). Deep Web Research and Discovery Resources 2012. *Virtual Private Library*.

(14) URL Shortening. (2012, November 23). In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/URL_shortening

(15) Sebastian. (2008, October 20). Crawling Vs Indexing. Retrieved from http://sebastians-pamphlets.com/crawling-vs-indexing/

(16) Web Search Engine. (2012, November 25). In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Web_search_engine

(17) Open Source Search Engine Software. (2012). In *SeekQuarry*. Retrieved from http://www.seekquarry.com

Thank you