RAID ON CODE PIRATE

- A Plagiarism Detection System

Supervisor
Mr. Daya Sagar Baral

Project Members
Kailash Budhathoki
Rakesh Manandhar
Shilpa Singhal

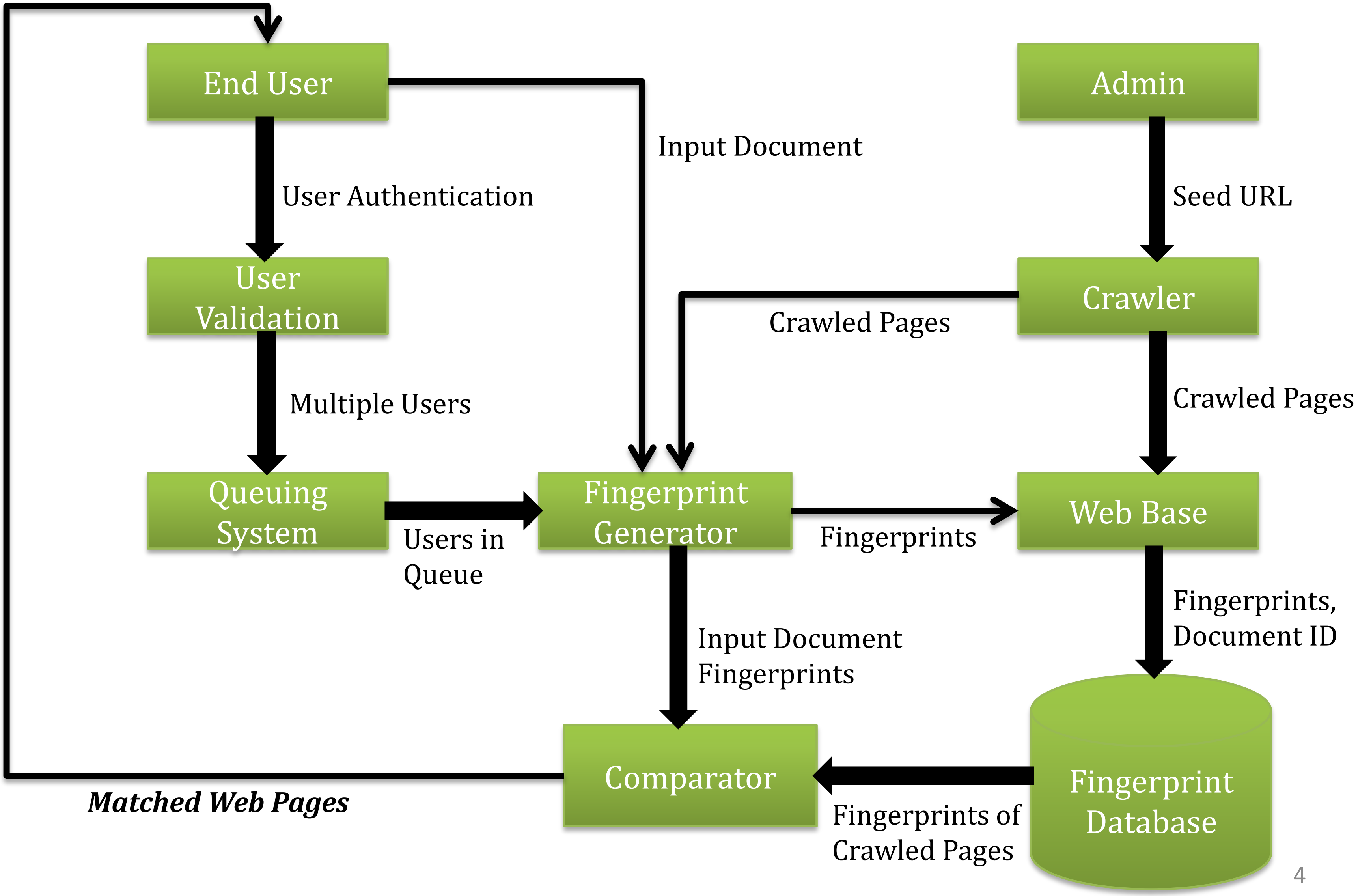
Introduction

- What is plagiarism?
  - Using other's ideas, thoughts, work without acknowledging the source of that information
- Detects the plagiarism in plain texts and source codes
- Implements structure metric detection technique
- Web based application
- Client-Server Architecture

Objectives

- To develop a web crawler capable of crawling the web pages under the same domain
- To create a web base of size 10 MB containing pages of shortlisted sites
- To develop a program that checks the provided documents with the pages in web base within the time constraint imposed for plagiarism

System Architecture



System Components

- Preprocessing of the input document
  - Removes irrelevant features(whitespaces, cases, etc)

    A do run run run a do Run run

    adorunrunrunadorunrun
- Fingerprint generator
  - Generates fingerprints
  - 3 steps
    - Generation of k-grams
      - K-grams = Contiguous substring of length k

    adoru dorun orunr runru unrun nrunr runru unrun nruna runad unado nador adoru dorun orunr runru unrun

System Components (Contd …)

- Generation of Hash Values
  - Uses Karp-Rabin rolling hash function
  - Sample Hash Value Calculation

K-gram = ‘adoru’

ASCII Value for ‘a’ = 97, ‘d’ = 100, ‘o’ = 111, ‘r’ = 114, ‘u’ = 117

Hash Value =

$97*101^{4}+100*101^{3}+111*101^{2}+114*101^{1}+117*101^{0}$

77 74 42 17 98 50 17 98 8 88 67 39 77 74 42 17 98

System Components (Contd …)

77 74 42 17 98 50 17 98 8 88 67 39 77 74 42 17 98

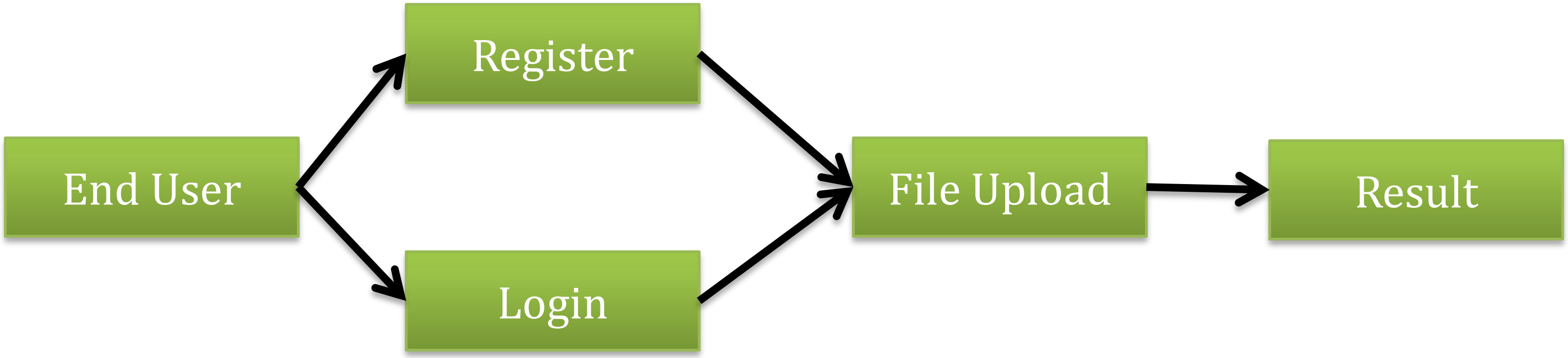
- Winnowing
  - Windows of hashes of length 4

[77 74 42 17] **[74 42 17 98]**
[42 17 98 50] ***[17 98 50 17]***
[98 50 17 98] **[50 17 98 8]**
[17 98 8 88] **[98 8 88 67]**
[8 88 67 39] **[88 67 39 77]**
[67 39 77 74] **[39 77 74 42]**
[77 74 42 17] **[74 42 17 98]**

17 17 8 39 17 ⟶ Fingerprints

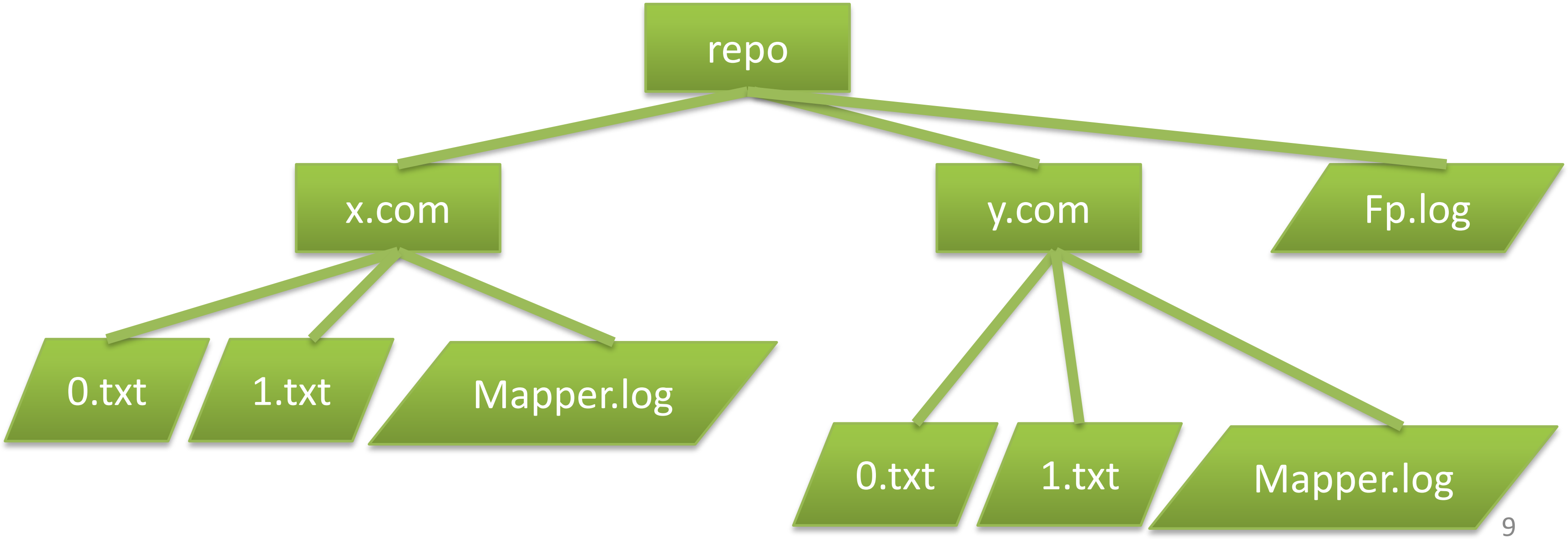
System Components (Contd …)

- Fingerprint comparator
  - Queries each fingerprint against the database
- Graphical User Interface
  - Web front end
  - Built using Django framework



System Components (Contd …)

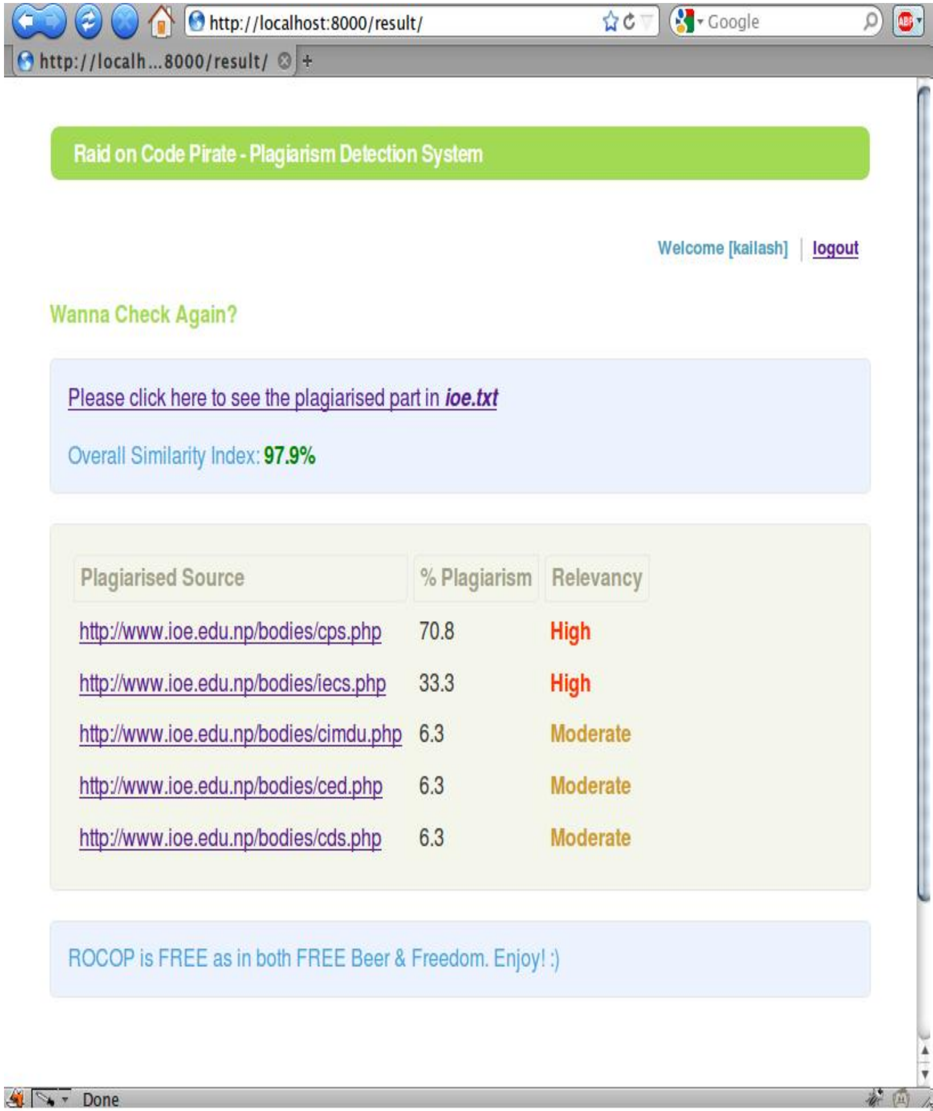
- Web Base creator
  - Updates the local repository
- Fingerprint database maintainer
  - Maintains a log file containing the list of websites whose fingerprint are already on the DB



Project Tools

- Platform: Ubuntu
- Programming Language: Python
- Web Framework: Django
- Third Party Library: Chilkat
- Database: MySQL
- Testing: PyUnit
- Tracking: D2Labs
- Versioning: SVN

Output



Raid on Code Pirate - Plagiarism Detection System

Welcome [kailash] | logout

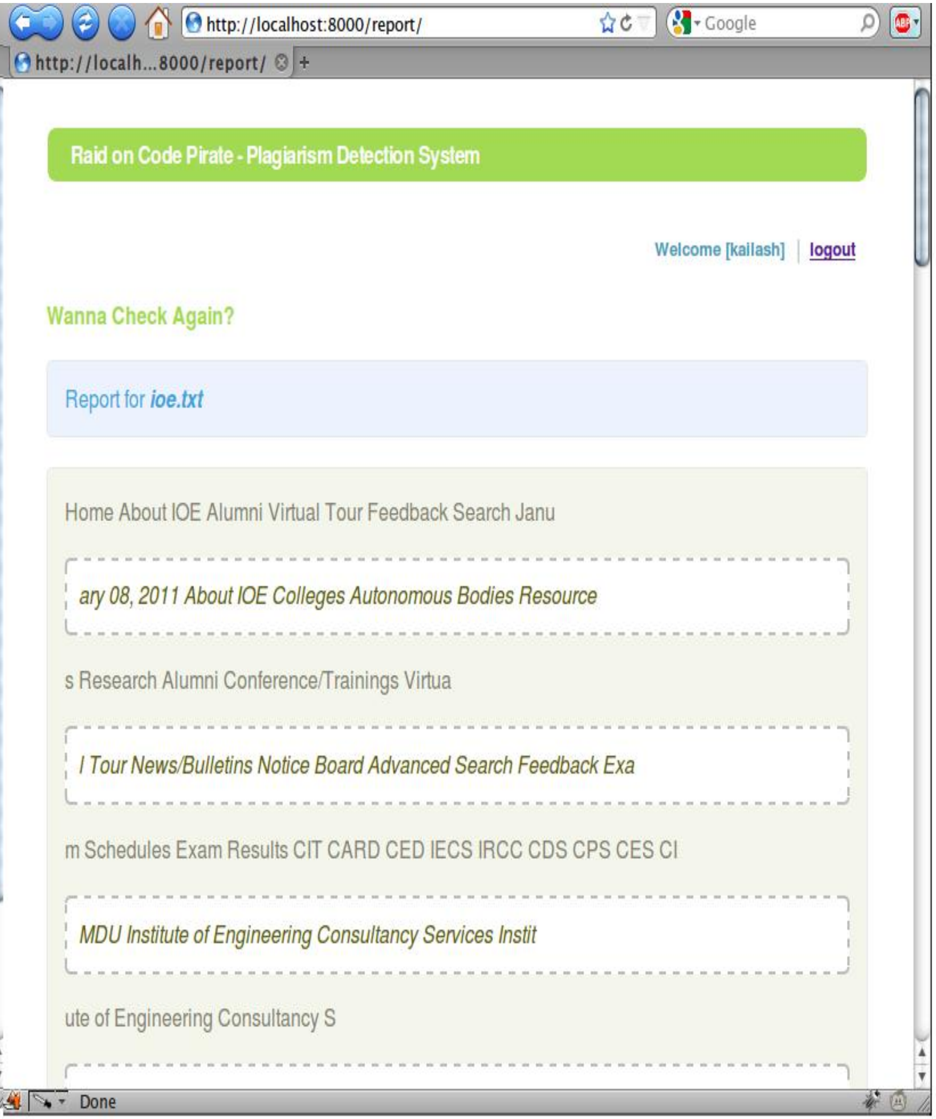
Wanna Check Again?

Please click here to see the plagiarised part in *ioe.txt*

Overall Similarity Index: **97.9%**

Plagiarised Source	% Plagiarism	Relevancy
http://www.ioe.edu.np/bodies/cps.php	70.8	High
http://www.ioe.edu.np/bodies/iecs.php	33.3	High
http://www.ioe.edu.np/bodies/cimdu.php	6.3	Moderate
http://www.ioe.edu.np/bodies/ced.php	6.3	Moderate
http://www.ioe.edu.np/bodies/cds.php	6.3	Moderate

ROCOP is FREE as in both FREE Beer & Freedom. Enjoy! :)



Raid on Code Pirate - Plagiarism Detection System

Welcome [kailash] | logout

Wanna Check Again?

Report for *ioe.txt*

Home About IOE Alumni Virtual Tour Feedback Search Janu

ary 08, 2011 About IOE Colleges Autonomous Bodies Resource

s Research Alumni Conference/Trainings Virtua

l Tour News/Bulletins Notice Board Advanced Search Feedback Exa

m Schedules Exam Results CIT CARD CED IECS IRCC CDS CPS CES CI

MDU Institute of Engineering Consultancy Services Instit

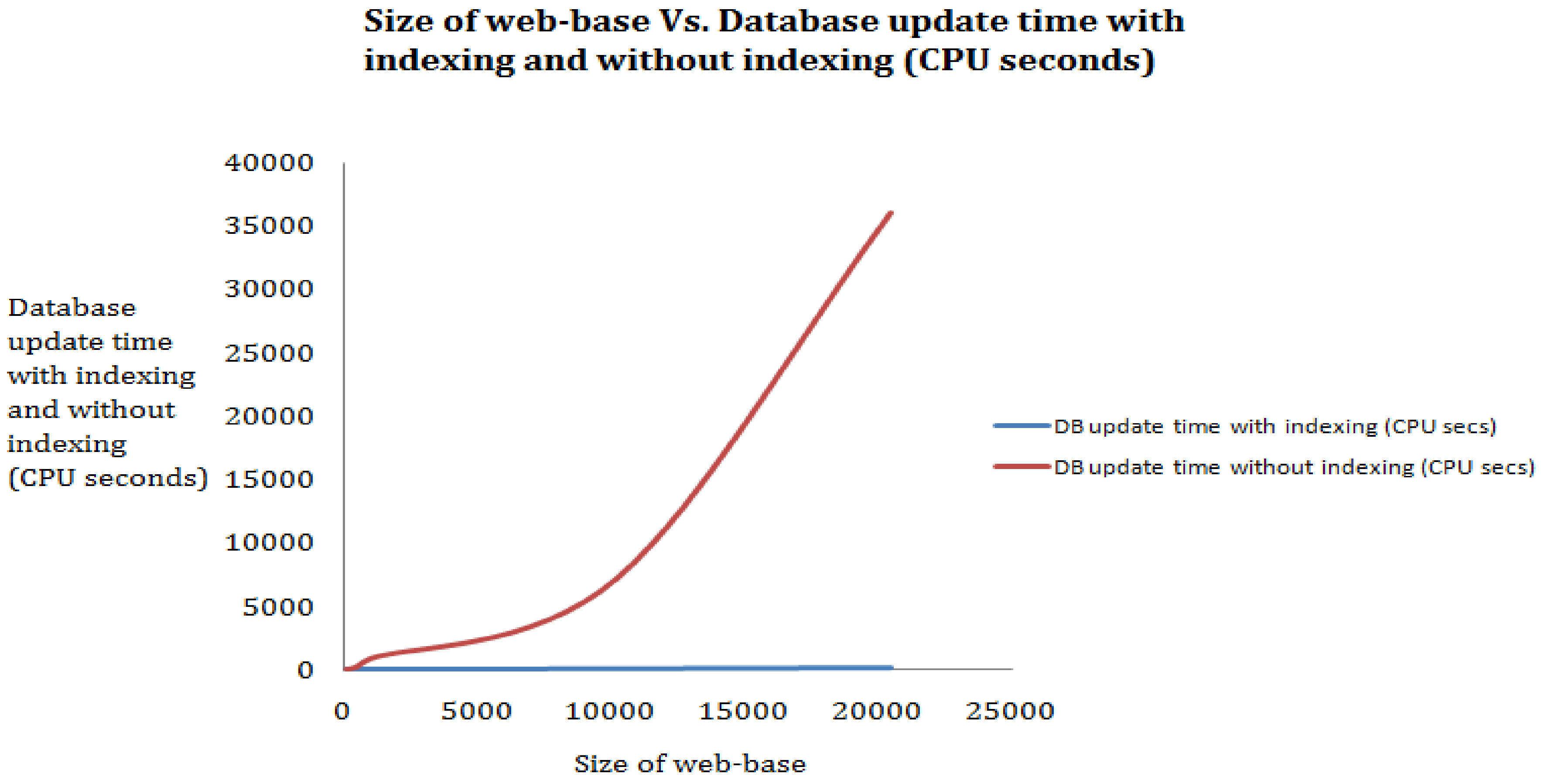
ute of Engineering Consultancy S

Comparison with Viper

S. No.	Features	ROCOP	Viper
1	Free/Open Source	Free and Open Source Software	Free (on monetary basis)
2	File Format	.txt	.doc, .pdf, .html, .rtf, .cs, .java
3	Client Interface	Web Page	Viper Client (software must be downloaded for use)
4	Platform Support	Platform independent	Windows only
5	Upload Limit	500 KB	Unlimited
6	Database Size	Small	Large (10bn resources)
7	Comparison Algorithms	Hashing, Winnowing	undisclosed
8	Detect Citation	No	Yes
9	Threshold	50 characters	No such threshold limit
10	Reliability	Higher	High
11	Analysis Time (for file size of 3KB)	1.87 seconds	3 seconds
12	Accuracy (for a particular document which is replicated from a page in the web-base)	97%	100%
13	Percentage similarity index	Yes	No
14	Links to plagiarized work	Yes	yes
15	Scope of search	Internal Database	Internal Database
16	Relevancy	Yes	Yes
17	Accepts an empty file	No	No

Optimization

- Indexing the table structure in database



Size of web-base Vs. Database update time with indexing and without indexing (CPU seconds)

Optimization (Contd…)

- Multi-processing Vs. Multi-threading
  - Scaling for multiple cores

- Different implementation of winnowing loop
  - Complexity issues

Application Area

- Implementation in colleges for detecting plagiarism in assignments submitted by students

Future Work

- Using NoSQL
- Implementing the system in distributed server architecture
- Using better algorithm to find the consecutive k-grams match
- Enhancing security measures (captcha)
- Using a distributed crawler
- Compressing the crawled content
- Fixing DB update issues
- Implementing the ability
  - To detect citation
  - To insert reference

We can no other answer make, but,
thanks, thanks and thanks.
~William Shakespeare