

# Title Based Duplicate Detection of Web Documents

Mrs. M. Kiruthika<sup>1</sup>, Mrs. Smita Dange<sup>2</sup>, Mrs. P. Sandhya<sup>3</sup>

<sup>1 2 3</sup>Department of Computer Engineering  
FCRIT, Vashi

<sup>1</sup>[Email- venkatr20032002@gmail.com](mailto:venkatr20032002@gmail.com), <sup>2</sup>[Email- smm\\_31@rediffmail.com](mailto:smm_31@rediffmail.com),

<sup>3</sup>[Email- sandhyachandranp@gmail.com](mailto:sandhyachandranp@gmail.com)

**Abstract-** In recent times, the concept of web crawling has received remarkable significance owing to extreme development of the World Wide Web. Very large amounts of web documents are swarming the web making the search engines less appropriate to the users. Among the vast number of web documents are many duplicates and near duplicates i.e. variants derived from the same original web document due to which additional overheads are created for search engines by which their performance and quality is significantly affected.

Web crawling research community has extensively recognized the need for detection of duplicate and near duplicate web pages. Providing the users with relevant results for their queries in the first page without duplicates and redundant results is a vital requisite. Also, this problem of duplication should be avoided to save storage as well as to improve search quality.

The near duplicate web pages are detected followed by the storage of crawled web pages in to repositories. The detection of near duplicates conserves network bandwidth, brings down storage cost and enhances the quality of search engines. In this paper, we have discussed a feasible method for detection of near-duplicate web documents based on the title of the documents which will help to reduce the overhead of search engines and improve their performance.

**Keywords** –Watermarking, Haar Wavelet, DWT, PSNR

## 1. INTRODUCTION

### 1.1 DETECTION OF NEAR DUPLICATE WEB DOCUMENTS-AN OVERVIEW

Exact Duplicate web documents are documents that do not differ from each other and they are exact copies of one another. But Near Duplicate documents are variants derived from the same original web document (i.e.) they are strikingly similar and they possess some minute difference and hence not regarded as exact duplicates.

Typographical errors, versioned, mirrored, or plagiarized documents, multiple representations of the same physical object, spam emails generated from the same template are some of the causes for near duplicate page generation. Such near duplicates vary only in minimal areas of the document like the advertisements, counters and timestamps and their content is similar. Web searches consider these differences as inappropriate.

Documents that are exact duplicates of each other are easy to identify by standard Check Summing techniques. But identification of near-duplicate documents is difficult.

## 2. EXISTING SYSTEM

The World Wide Web includes some document duplicated in different forms and at different places. Some documents are mirrored at different sites on the web. Some documents have different versions.

Some previous techniques for deleting duplicate and near duplicates document involve generating fingerprints. Two documents are considered to be near duplicates if they share more than a predetermined number of fingerprints. For large collection of documents this technique becomes expensive in terms of computation and storage.

Broder described system in which regions of each document called shingles are each treated as sequence of tokens and reduced to numerical representation. These are then converted to fingerprint using method described by robin.

Bloomfield has described an algorithm for detecting plagiarism which simply searches for matches of size or more successive words between two documents.

Another method for detecting near duplicate was proposed by Mankuet et al. It included demonstration of fingerprinting technique. Secondly an algorithm is represented for identifying existing f-bit-fingerprints that differ from given fingerprint is almost K bit positions for small k.

## **2.1 APPROACHES**

A technique for the estimation of the degree of similarity among pairs of documents was presented in 1997 by Broder et al, which was known as shingling, does not rely on any linguistic knowledge other than the ability to tokenize documents into a list of words, i.e. it is merely syntactic. In shingling, all word sequences of adjacent words are extracted. If two documents contain the same set of shingles they are considered equivalent and if their sets of shingles appreciably overlap, they are exceedingly similar.

An approach, based on similarity metrics for the detection of duplicated pages in web sites and applications, implemented with HTML language and ASP technology was proposed by Di Lucca et al.

Llyinsky et al. suggested a method of "descriptive words" for definition of near duplicate of documents, which was on the basis of the choice of N words from the index to determine a "signature" of a document. Any search engine based on the inverted index can apply this method. Similarly there are several other approaches proposed to solve the problem of duplication. But the need for various forms of duplicate document detection has increased due to accelerated growth of massive electronic data environments, both Web-based and proprietary. This detection can take any of several forms based on the nature of the domain and its customary search paradigms.

## **2.2 PROPOSED SYSTEM**

The growth of the internet challenges search engine as more copies of web documents flood over search results making them less relevant to users. The nature of copies is wide. The same document served from the same server may differ because of technical reasons like different character sets formats and inclusions of advertisement or current date. On the other hand similar documents are massively generated by database servers e.g. Messages in web forums, product pages in e-shops etc. In this paper, we would like to discuss a system which would perform near duplicate detection based on the title of the documents.

## **3. DESIGN**

### **3.1 MODULES**

Our system has the following modules with different functionalities.

#### **1] Database Module**

Database module would be a simulation of the World Wide Web. The database will contain the URLs related to certain topics. These URLs may direct us to some web documents that are variants of a single web document or to exact duplicate pages or even to some non-duplicate links. Basically the database will provide links.

#### **2] Search Module**

Search module will simulate a search engine. Search engines retrieve pages from the web server whereas this module will extract URLs from the database. Since our database contains links related to various topics it is necessary to search links related to user's keywords from the database.

#### **3] Extraction Module**

This module deals with detection of near duplicate web documents based on the title of the web document. Hence we need a module to extract the title of the document from its html code. Extraction module performs two tasks. Firstly it retrieves the html code of the web page and then extracts title from the code.

#### **4] Comparison and Detection Module**

The comparison and detection module helps to reach to the final result. Firstly it compares the title of web document with the user's keywords. Then it appropriately classifies the result of the match and leads us to the final output of whether duplicates are detected or not.

## Title Based Duplicate Detection of Web Documents

### 3.2 DESIGN STEPS

**INPUT:** Keywords entered by the user

**PROCESS:** Find URL from the database that are related to the keywords

1. Extract title from each URL
2. Compare the title with the keywords entered by the user
3. If (match found) Duplicates exist  
Else  
No duplication

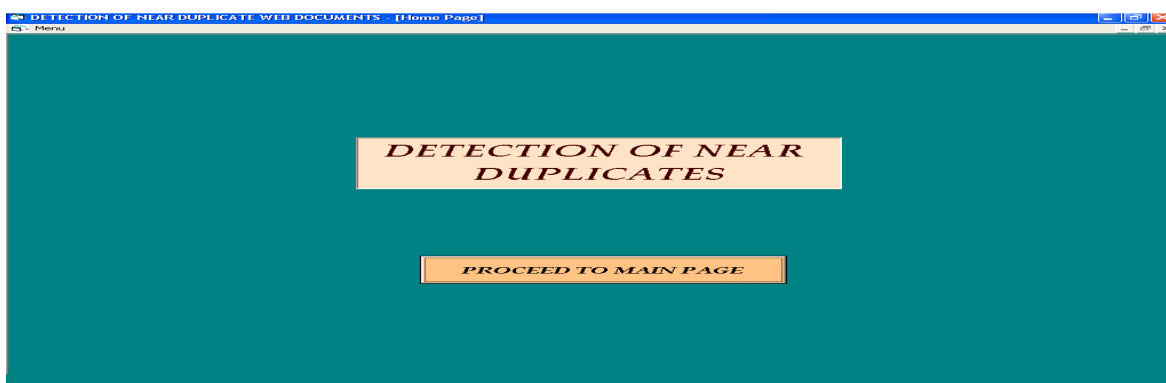
**OUTPUT:** Duplicate URLs

In our system, user enters the keywords to find the near duplicate links. Then system searches links related to keywords entered by the user and displays it. System asks user whether or not to detect near duplicate links. If user wants to detect near duplicates then system extracts the title of each URL from related links. If title matches with the user's keywords then match is found and near duplicate links are displayed.

## 4 IMPLEMENTATION

The system has been implemented and the following screenshots will help to understand the practical implementation of each module and of the overall system.

### 1] Home page:



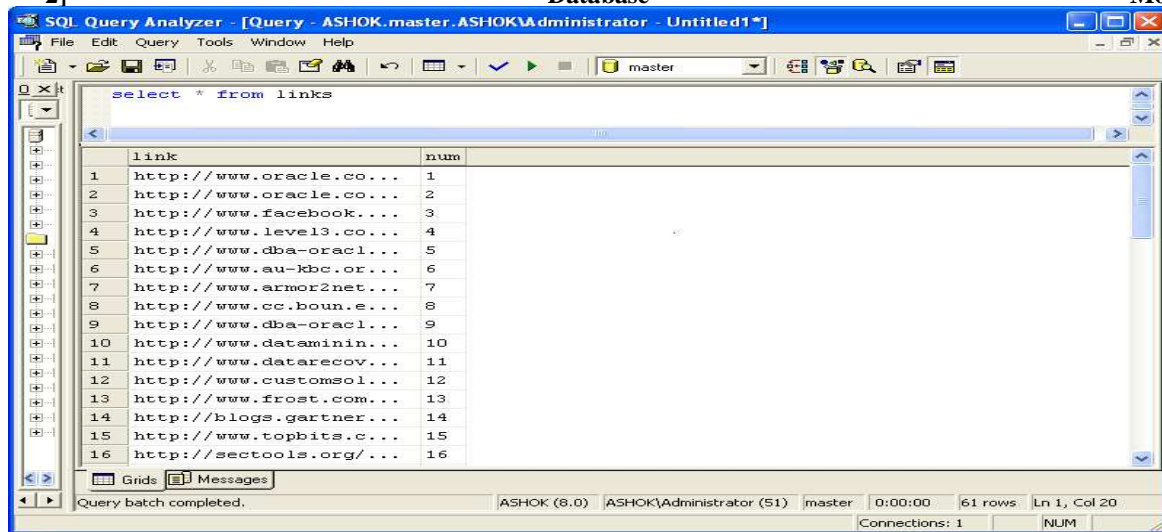
This is the first screen of our system. This page provides an option Menu in which two sub options are provided

1. Detection of pages- for proceeding to detect near duplicate web pages.
2. Exit-for exiting the system.

2]

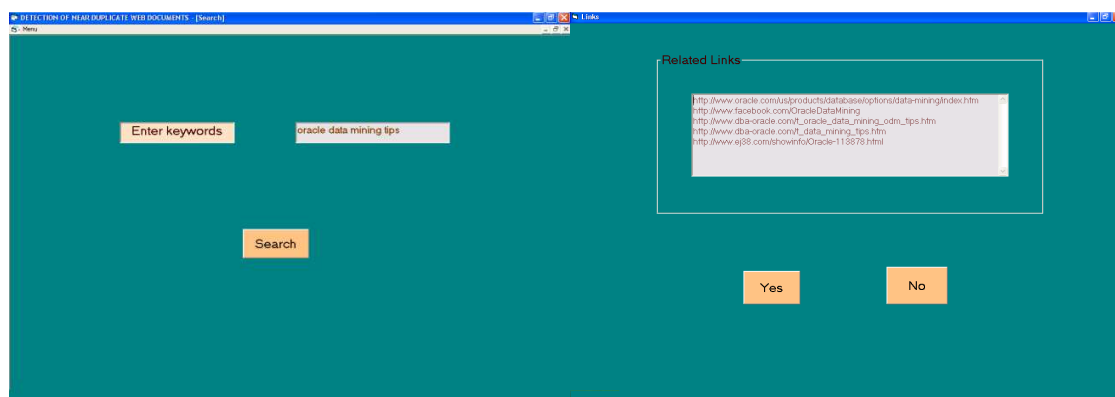
## Database

Module:



Database module is implemented using SQL. The above screenshot shows some of the URLs from our database. Some of these URLs direct us to some web documents that are variants of a single web document or to exact duplicate pages or even to some totally different pages.

### 3] Search Module:



Search module is used to find links from the database that are related to user's keywords. In our system we ask the user to first enter the keywords and the GUI for that is shown in the first screenshot. On clicking the Search button the module is activated and connection is established with the database. A search is performed to find links related to the keywords and the related links from the database are displayed. If for certain keywords match is not found then an appropriate message is given to the users. The second screenshot depicts the results of a search on the keywords "oracle data mining tips".

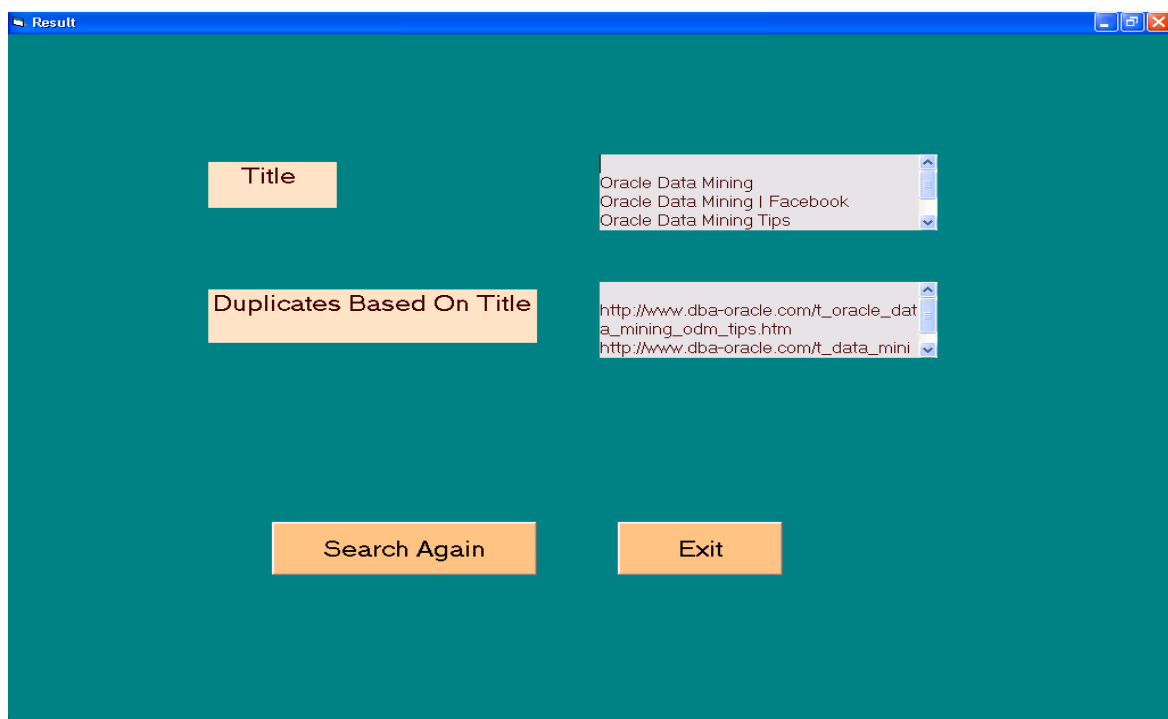
### 4] Extraction Module:

## Title Based Duplicate Detection of Web Documents



The input to this module is the list of URLs from the related links(output of the search module).This module first sends an http request for each URL. The http request returns the html code of the page. From the html code the module extracts the title and that is displayed in the result. The above screenshot displays the title of each URL in the first Result box named TITLE

### 5]Comparison and Detection Module:



This module is used to compare the titles extracted from the URLs with the keywords entered by the user. The result of comparison is determined in terms of number of matches. There can be three cases for the number of matches.

**Case 1:**

If the number of matches is Zero that means none of the title matched with the user's keywords and hence no question of duplication.

**Case 2:**

If the number of matches is One that means only one link matches with user's keywords and hence there is no duplication. For this case the module displays the link found and also tells the user that there is no duplication.

**Case 3:**

If the number of matches is two or greater than two that means near duplicate web pages exist for the keywords entered by the user. Hence the module displays the links which are leading to duplicate documents. This case is depicted in the screenshot in the result box titled " Duplicates Based On Title"

## 6. TESTING

### 6.1 TEST CASES

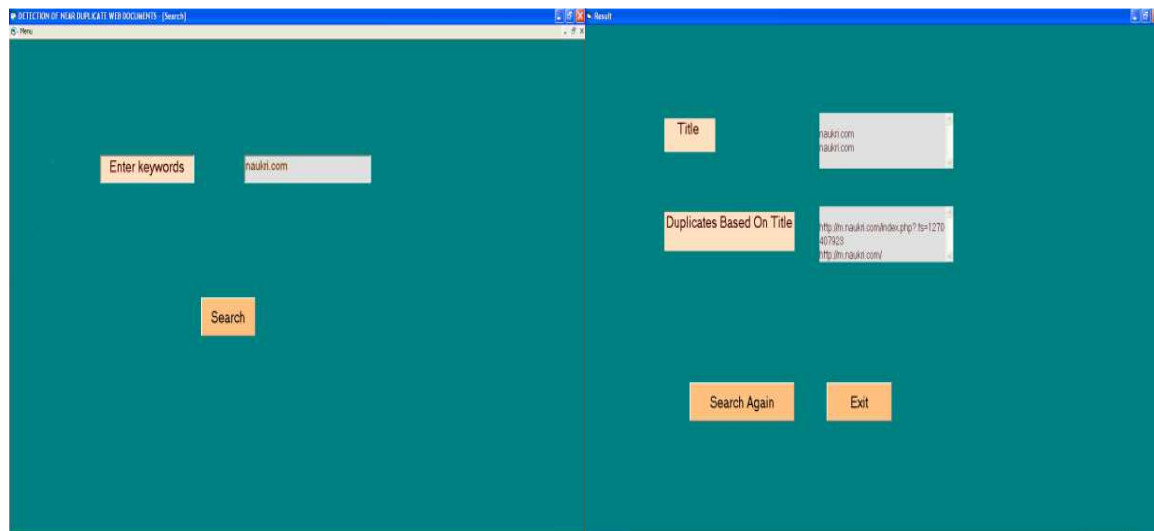
Sr. No.	Test case	Expected Output	Actual Output	Remark
1	naukri.com	http://m.naukri.com/index.php?.ts=1270407923 http://m.naukri.com/	http://m.naukri.com/index.php?.ts=127040793 http://m.naukri.com/	Near duplicate links detected.
2	security	There are links in database but none have the title that matches user's keywords	System displays message(NO MATCH FOUND)	No match found
3	network security	Since there are no duplicate links system should give a message-'only one match found'.	System displays the link and gives message (NOTE:ONLY ONE MATCH FOUND).	No near duplicate links found.
4	Image processing	Since no links of this topic are stored in the database, system should give message asking the user to enter other keywords	No related link found in database. Please enter some other keywords.	Ok.

**Test case 1:**

naukri.com

User enters keyword

## Title Based Duplicate Detection of Web Documents

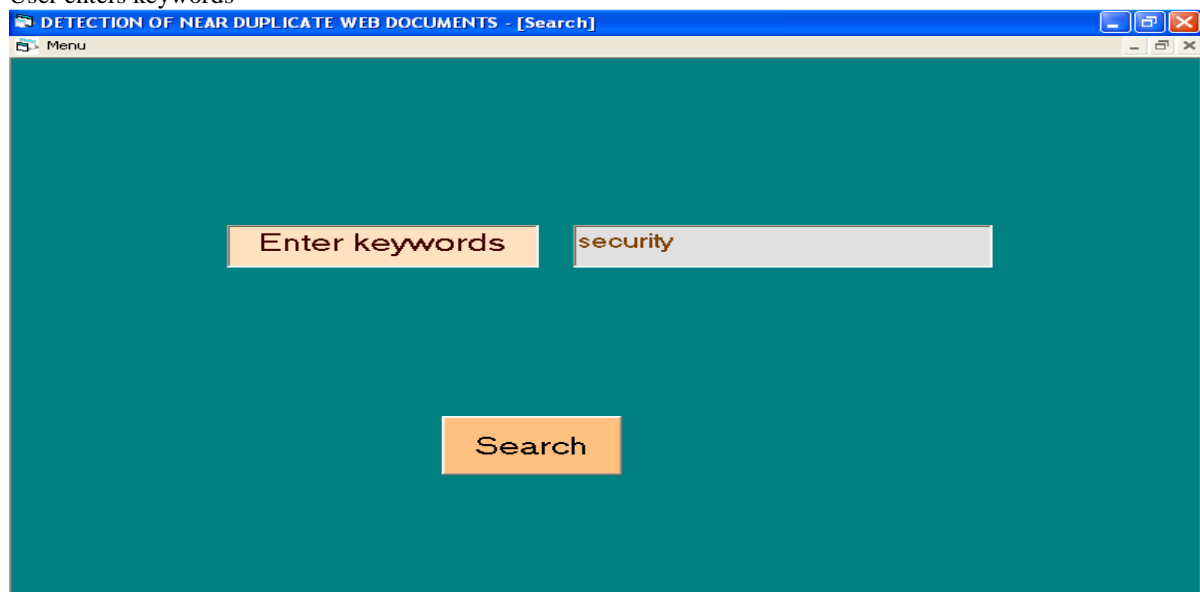


Near duplicate links detected.

### Test Case 2:

security

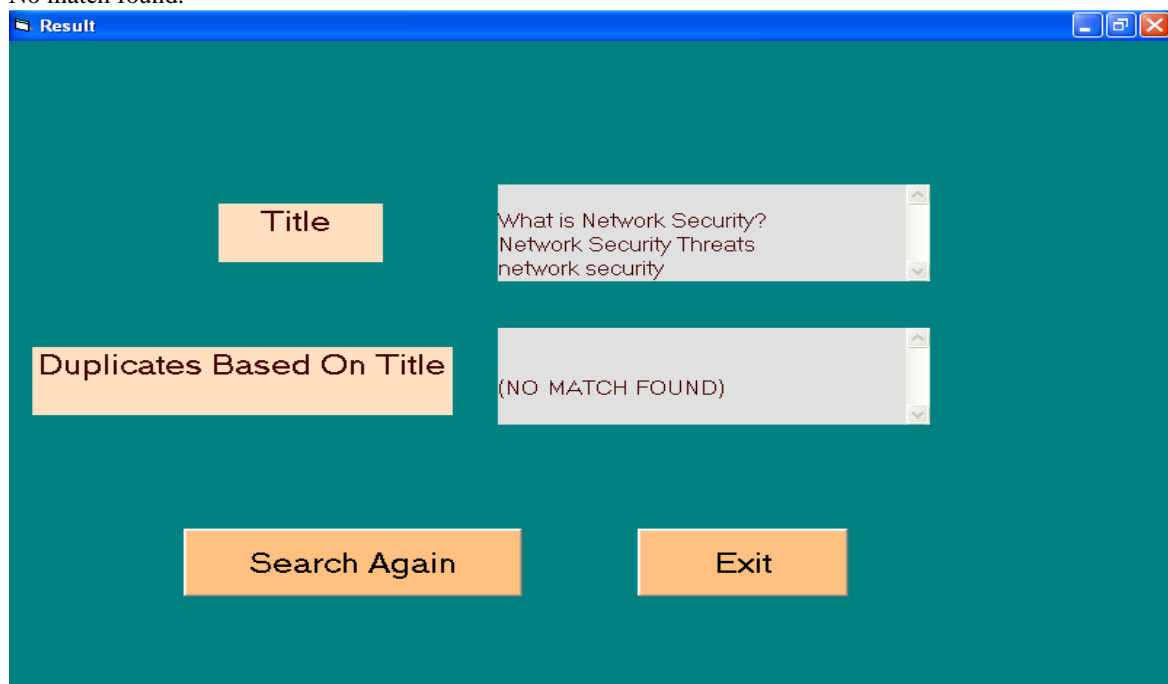
User enters keywords



RELATED LINKS ARE DISPLAYED



No match found.



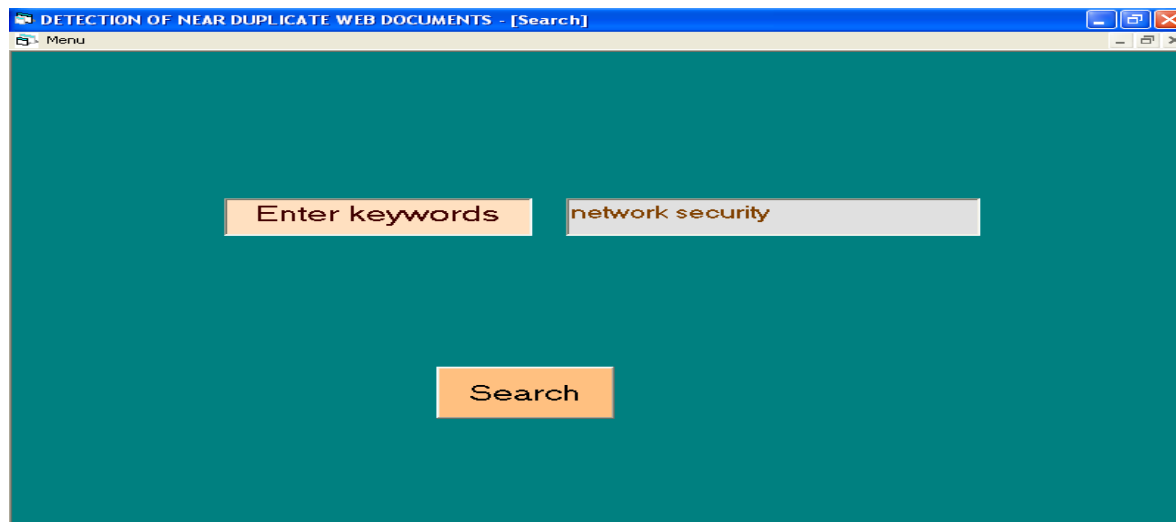


## Title Based Duplicate Detection of Web Documents

### Test case 3:

Network security

When user enters keyword.



DETECTION OF NEAR DUPLICATE WEB DOCUMENTS - [Search]

Menu

Enter keywords

network security

Search

Related links are displayed.



Links

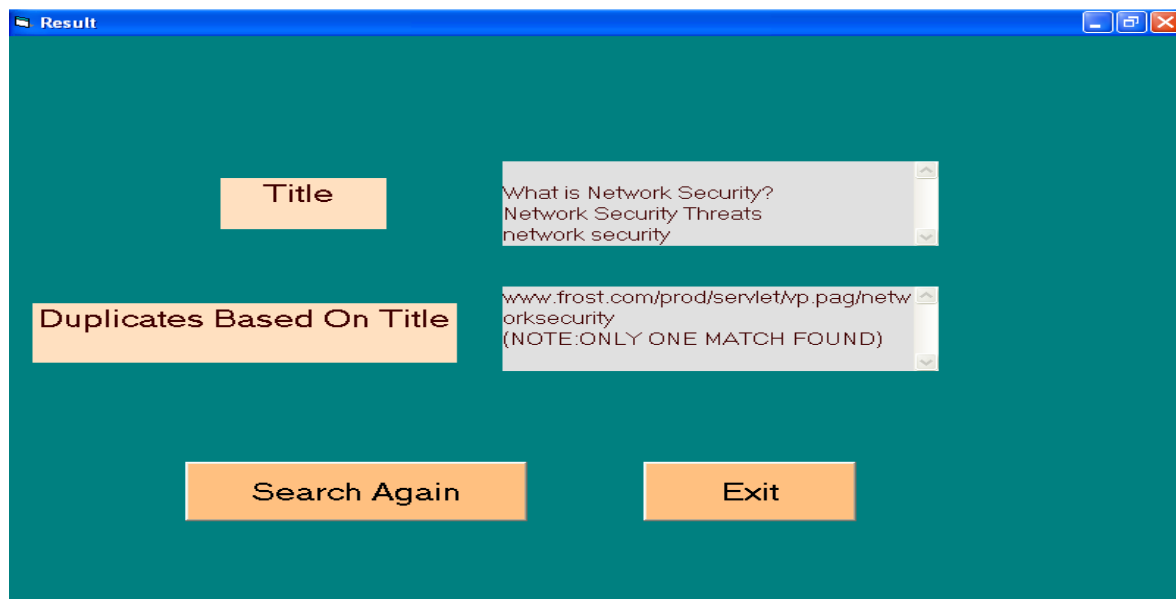
Related Links

<http://www.wisegEEK.com/what-is-network-security.htm>  
<http://www.spamlaws.com/network-security-threat.html>  
<http://www.frost.com/prod/servlet/vp.pag/networksecurity>

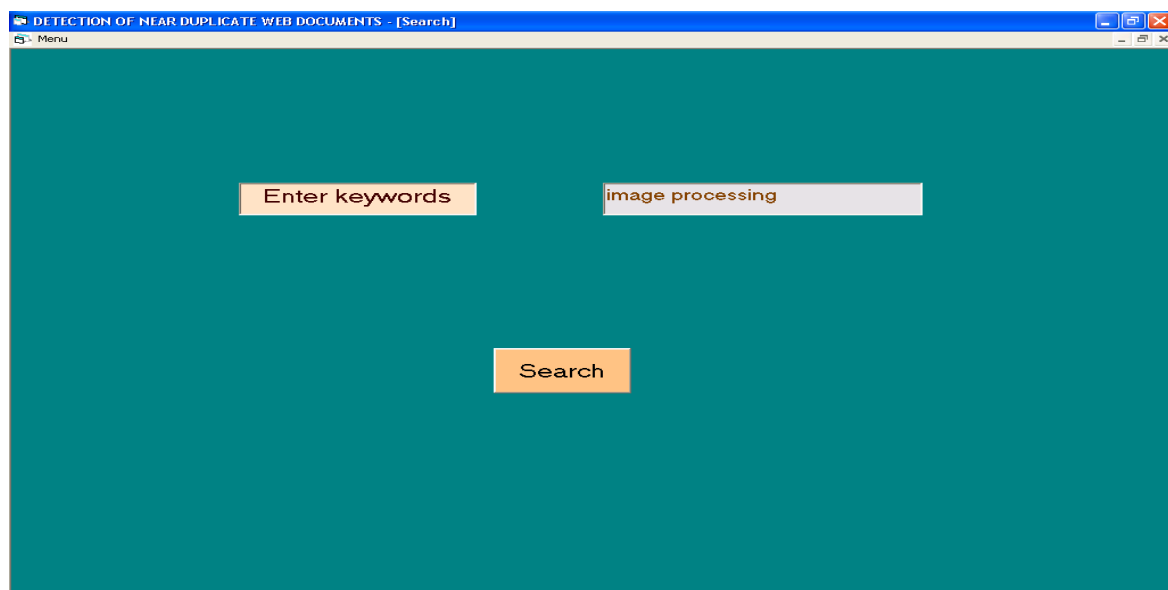
Yes

No

Only one match found.



**Test case 4:**  
Image processing  
When user enters keyword.



## 8. CONCLUSION AND FUTURE SCOPE

## Title Based Duplicate Detection of Web Documents

The developed system detects duplicates based on the title from set of URLs that are stored in the database. Therefore, to some extent the system is viewed as static and the performance accuracy was reasonably good. There are several aspects which have to be addressed. The further possible improvements for system are: -

- It can be enhanced to extract duplicates dynamically.
- We can extract author, content, file format and compare the documents to detect near duplicates with more efficiency.

## REFERENCES

- [1] J. Cho, H. Garcia-Molina, and L. Page, "Efficient crawling through URL ordering", Computer Networks and ISDN Systems, vol. 30, no. 1-7: pp. 161-172, 1998.
- [2] M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused crawling using context graphs", In 26th International Conference on Very Large Databases, (VLDB 2000), pages 527-534, Sep 2000.
- [3] F. Menczer, G. Pant, P. Srinivasan, and M. E. Ruiz, "Evaluating topic-driven web crawlers", In Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 241-249, 2001.
- [4] S. Pandey and C. Olston, "User-centric web crawling", In Proc. WWW 2005, pages 401- 411, 2005.
- [5] Gurmeet Singh Manku, Arvind Jain, Anish Das Sarma, "Detecting near-duplicates for webcrawling", Proceedings of the 16th international conference on World Wide Web pp. 141 - 150, 2007.
- [6] Fetterly, D. Manasse, M. Najork, M., "On the evolution of clusters of near-duplicate Web pages", Proceedings. First Latin American Web Congress, 10-12 Nov. 2003, pp.37- 45
- [7] Dennis Fetterly, Mark Manasse, Marc Najork, "Detecting phrase-level duplication on the world wide web", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp.170 - 177, 2005.
- [8] Hui Yang, Jamie Callan, "Near-duplicate detection for eRulemaking", ACM International Conference Proceeding Series; Vol. 89, Proceedings of the 2005 national conference on Digital government research, pp.78 - 86, 2005.
- [9] Hui Yang, Jamie Callan, Stuart Shulman, "Next steps in near-duplicate detection for eRulemaking", Proceedings of the 2006 national conference on Digital government research, May 15-18, 2006, pp. 239 - 248.
- [10] Ee-Peng Lim and Aixin Sun, "Web Mining - The Ontology Approach ", 2005.
- [11] Ziv Bar-Yossef, Idit Keidar, Uri Schonfeld, "Do not crawl in the dust: different urls with similar text," Proceedings of the 16th international conference on World Wide Web, pp: 111- 120, 2007.
- [12] Chuan Xiao, Wei Wang, Xuemin Lin, Jeffrey Xu Yu , "Efficient Similarity Joins for Near Duplicate Detection", Proceeding of the 17th international conference on World Wide Web, pp:131--140, 2008.
- [13] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic clustering of the web", In Proceedings of Sixth International Conference on World Wide Web, pp: 1157-1166, 1997.
- [14] Jack G. Conrad, Xi S. Guo, and Cindy P. Schriber "Online Duplicate Document Detection: Signature Reliability in a Dynamic Retrieval Environment," In Proceedings of the 2003 ACMCIKM .Twelfth International Conference on Information and Knowledge Management (CIKM03) (New Orleans, Louisiana), ACM Press, New York, pp. 243-252, 2003.
- [15] D. Metzler, Y. Bernstein and W. Bruce Croft. "Similarity Measures for Tracking Information Flow", Proceedings of the fourteenth international conference on Information and knowledge management, CIKM'05, October 31-November 5, 2005, Bremen, Germany.