Detecting a Broken Link in a Web Site

Dr. JAMAL F. TAWFEQ* Dr. ABDUL MONEM S. RAHMA** EHSAN Q. AHMED*

Abstract

Broken links in a Web site is common problem on the Internet today. It inconvenience visitors, inhibits proper navigation of a site, prohibit access to Web site content and reduce the productivity of Web professionals. Furthermore, as our society becomes more dependent upon the Internet, ensuring that links are accurately, efficiently and timely maintained will assume a heightened priority.

The aim of this paper is to find a method that test all website pages links, and gives a detailed report about all found broken links, whether these links were text links or image links.

*Al-Nahrain University
** University Of Technology

Introduction

One of most serious problems plaguing the World Wide Web (WWW) today is that of broken hypertext links, which are a major annoyance to browsing users and also a cause of tarnished reputation and possible loss of opportunity for information providers. The root of the problem lies in the current Web architecture's lack of support for referential integrity{1}.

The WWW is also a distributed hypermedia environment consisting of documents from around the world. The documents are linked using a system known as Hypertext, where elements of one document may be linked to specific elements of another document. The documents may locate on any computer connected to the internet. In this context, the world "document" is not limited to text but may include video, audio, graphics, databases, and a host of other tools that can be accessed from any web browser {2}.

These documents are created with a special language called Hypertext Markup Language(HTML).This language allows the full use of the hypermedia including text, images, graphics, sounds and other types of multimedia. Because HTML is a special language it requires special software to access the web. This type of access program is known as Browser **{3}.**

Web Page

A Web page is a resource of information that is suitable for the WWW and can be accessed through a web browser. This information is usually in HTML or XHTML format, and may provide navigation to other web pages via hypertext links.

Web pages may be retrieved from a local computer or from a remote web server. The web server may restrict access only to a private network, e.g. a corporate intranet, or it may publish pages on the

WWW. Web pages are requested and served from web servers using Hypertext Transfer Protocol (HTTP).

Each HTML document you create is a single web page, regardless of the length of the document or the amount of information included {4}.

Web Site

Web site is a collection of web pages under the control of a particular person or group. Generally, a web site offers a certain amount of organization of its internal information. The user might start with an index or default page for a web site and then use hypertext links to access more detailed information. Another page within the web site might offer links to other interesting sites on the web, information about the organization, or just about anything else {5}.

Any web site is represents by directory structure. Directories ("places" to store files) are organized into a hierarchical structure that fans out like an upside-down tree. The top-most directory is known as the *root* and is written as a forward slash (/). The root can contain several directories, each of which can contain subdirectories; each of these can contain more subdirectories, and so on. A subdirectory is said to be the "child" of the directory that holds its "parent". Figure (1) shows a system with five directories under the root. The directory *users* have two subdirectories, *jen* and *richard*. Within *jen* are two more subdirectories, *work* and *pers*, and within *pers* is the file art.html{5}.

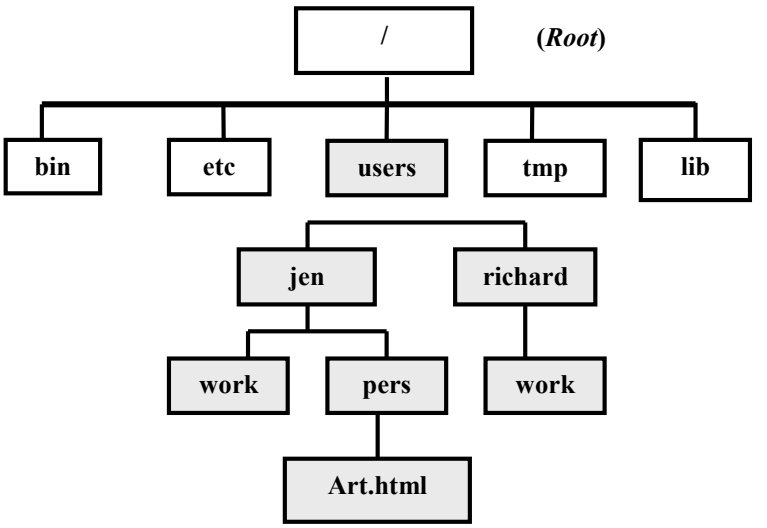

Figure 1
The Strucutre of Web Site

Uniform Resource Locator (URL)

The URL is a string of characters that represents the location or address of a resource on the Internet and how that resource should be accessed. World Wide Web pages are assigned a unique URL. Each hyperlink on a web page contains the URL of the page to be linked to{6}.

Every document on the World Wide Web has a unique address. The document's address is known as its *uniform resource locator* (URL).

A URL consists of the document's name preceded by the hierarchy of directory names in which the file is stored (*pathname*), the Internet domain name of the server that hosts the file and the software and manner by which the browser and the document's host server communicate to exchange the document (protocol):

Protocol://server_domain_name/pathname

Types of Hyperlinks

Hyperlinks, also known simply as links, provide a means of cross referencing points within and across documents. They often appear as underlined or otherwise highlighted, text but can also be found in pictures. There are three types of HTML hyperlinks, and each one is used in different situation see figure (2){5}.

Absolute URLs: links to a page on a different web server.

Relative URLs: links to a page on the same web server.

Linking within a document: links to a different location on same web page.

Linking to fragment.

Linking to fragment in another document.

Figure (2) illustrates the difference between relative and absolute URLs.

If the user creates a web site of any complexity, he (her) will need all three types of hyperlinks{7}.

All hyperlinks have two components:

A link label (a clickable element on a web page).

A link destination (a target destination).

The link label and the link destination are both components found inside the special HTML element used to create hyperlinks. Here's example preview of a link in HTML:

cs120 homepage

Link destination **Link label**

1. The Broken Link's Problem

In a website structure, navigation problem raised due to broken links. The broken link may involve at various levels of web site structure{8}.

A dead link or broken link is a link on the World Wide Web that points to a web page or server that is permanently unavailable.

Most of internet's users faced the problem "file not found" in the internet which represent by error 404 this error message indicates that the server name is valid,

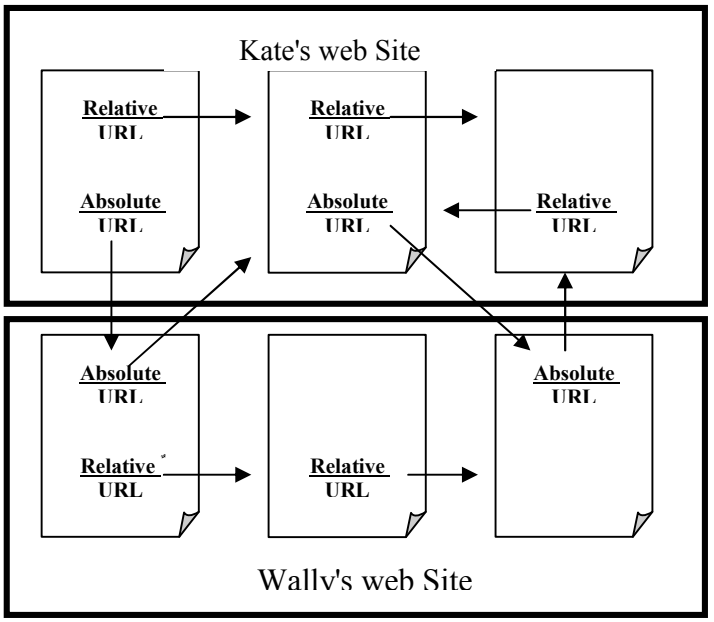

Figure (2)

Relative and Absolute URLs

but the web page could not be found at the specified location. Four things could be wrong{9}:

a. The web page might have moved. In this case, the user can try to find its new location.

b. Error in the syntax of link.

c. Changed the name of the web page.

d. Deleted the web page from the web site.

Nonworking links can frustrate web site visitors, so keep web page in good operating condition by periodically verifying that all links still work correctly. It is not enough to know that a link was working when first created it.

After all, a link that works today might not work tomorrow. The page might be removed from the web or the page's author might rearrange some files of directories, rendering the old URL obsolete{10}.

Ongoing maintenance is needed to ensure that hyperlinks remain operational next week, next month, and next year. This requirement is one of the hidden costs associated with posting pages on the web{11}.

2. Disadvantages of Broken Link's Problem

One of the most important indicators of a high-quality site is absence of broken links {12}.

a. There is nothing worse for a user than coming across either a link that leads nowhere or an empty square instead of an image or video.

b. Broken links ruin web site reputation and bring down its rating on search engines.

c. Search engine rates the web site lower if it finds broken links there.

Broken links are not just errors in the design of a site but they can cost rating traffic and money.

A Web page dies every time that one or more files on a Web server have their names changed, their location in the subdirectory structure moved, or their host names modified. Links also will die when a server name changes.

3. Proposal system for detecting broking link system architecture

The aim of this paper is to design software that test all website pages links, and gives a detailed report about all found broken links, whether these links were text links or image links. The structure of proposed link checker system consists of three main modules are crawler, extractor, and checker module as shown in Figure (3) each module has specific functions.

Broken links are counted at various levels and broken error index is calculated based on percentage of broken links involved in sitemap tree. The percentage of broken links is calculated using equation (1). It is represent the web site quality assessment.

Percentage of Bad links =

(number of bad links / number of web pages)*100 . . . (1)

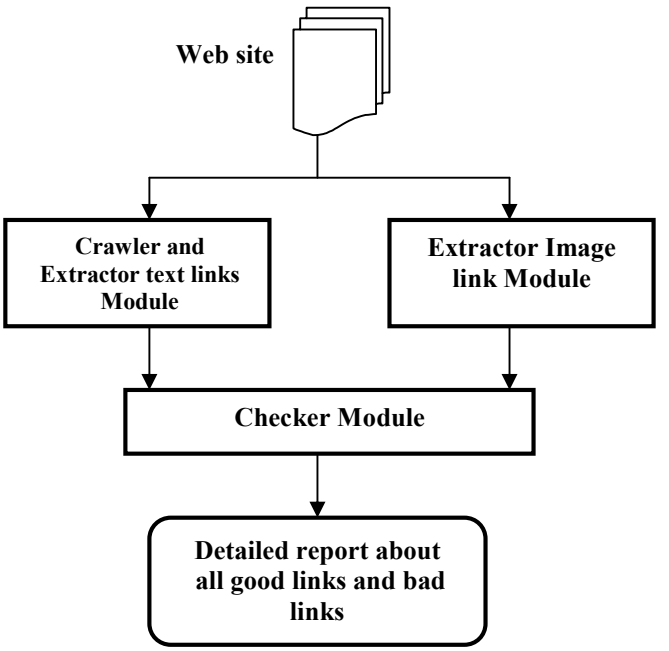

Figure 3
Proposal system for detecting broking link system architecture

8.1 Crawler and Extractor Module:

The crawling is a process to collect all website pages. This module will access all the pages of website starting with home page and then extract

all the text links founded within web pages, then classified according to their types (absolute link, relative link, fragment link). The crawler consists of two parts in this proposed link checker; as shown in Figure (4).

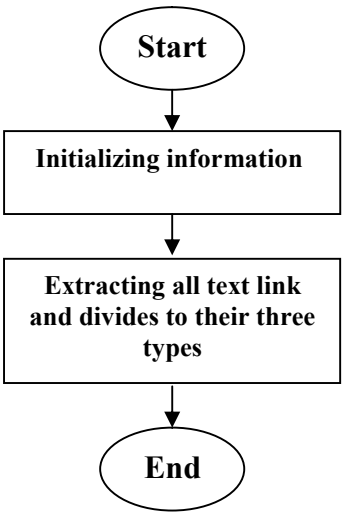

Figure 4
Crawler Module

These two parts are:

a. *Initializing the crawling information*: All information related to the crawling process is initialized, this information represents by this module will access all the pages of website starting with home page and then extract all the text links founded within web pages, then classified according to their types (absolute link, relative link, fragment link).

b. *Extracting links and classified them to their types*: In this step the crawling fetch links one by one from relative list then open it as file for reading to extracting all text links and classified to their three types (relative list, absolute list and fragment list).

After initializing relative-list with the name of home page, now fetching links from relative-list one by one begins with home page, after each page is checked, all links in that page are extracted out and then classified them to their types (i.e. if the link is relative link put it in

relative-list, if the link is fragment link put it in fragment-list, if the link is absolute link put it in absolute-list).

The crawler finishes its work when there is no more links in relative link list and the current link point to the last link(see algorithm(1)).

Algorithm(1)
crawling web pages

```
  Goal: Traverse relative-list starting with home page
  Input: all pages of web site
Output: relative-list, absolute-list and fragment list
Variables: link, fullpath as string

While (current-link less than or equal size-relative) do
      Set link ← relative-list (current-link)of link field
      Set fullpath-link execute concat-root(link)
      Execute search-page(link)          //To check If it is found in
                                                website or
                                                not

If search-page(link) return false value then
      Set relative-list (current-link) of found field ← false)
   Else
      Set relative-list (current-link) of found field← true)
End If

If  relative-list(current-link) of check field equal false then
      Execute read-page(link)          // To extract all links and
                                             classified to
                                             their types

Set relative-list(current-link)of check field ← true
End If
```

8.2 Extractor image link module:

All image links in web pages are extracted in this module**.** The proposed link checker system also traversing and checking the image links after traverse all three types of text links(relative-links, absolute links and fragment links).The system extracted an image link from relative link list because this list contains all links in web site.

There are number of steps will be done to extract image link. These steps are:

Step 1: at first of all open the page which is send as input parameter from algorithm (3.10) as file for reading.

Step 2: read from file character by character until smaller than "<" character will be found.

Step 3: reading process continuous if the word which comes after "<" character it is "img" word, then check the next word if it is "scr" an algorithm reach to an image link.

Step 4: after extracting image link, this link will be added to image link list but before this process the algorithm check if this image link added before this time to prevent duplication in the image link list, addition process also change the value of image link list fields.

8.3 Checker module:

This module consists of two sub-modules:

a. ***Checker text link sub-module***: All fragment links and absolute links will be checked. After isolate fragment name from link string, now using check fragment algorithm to traverse fragment links. In

this module the extracted links from read-page algorithm will be checked. Relative links checking while crawling executing.

There are two steps to check-fragment will be done:

Step 1: fetch links one by one from fragment list, then send link and destination page to traverse-fragment algorithm to check if fragment link found or not.

Step 2: if traverse-fragment algorithm return true value then the found field in fragment list becomes true and fetch another fragment link, stop condition of check-fragment algorithm is no more link in fragment list.

b. Checker image link sub-module: This model check-image used to checking if image exist in web site or not. All extracted image links will be checked.

4. Proposal system for detecting broking link system interface:

The proposed link checker system checking all links in website by receiving website's URL as input to extract all links and then classified to their types (relative links, fragment links, absolute links and image links).

In the main menu interface appear, the user select one website from the list of websites URLs which found in "Enter URL" field. So this interface considered as important one because of this selection the user can start website checking. This interface contains number of commands which explains in details in next section. (See Figure (5))

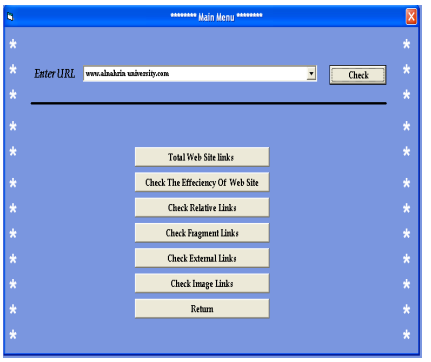

Figure 5
Main menu interface

When the user click on command button "Check Relative Links" the proposed system opened new interface which called "Relative Links Checking" as shown in figure (6)

Figure 6
Relative links checking interface

When the user click on Check Relative Links Command Button; the system display table of links information. Figure (7) show all relative links and the status of those (good or bad).

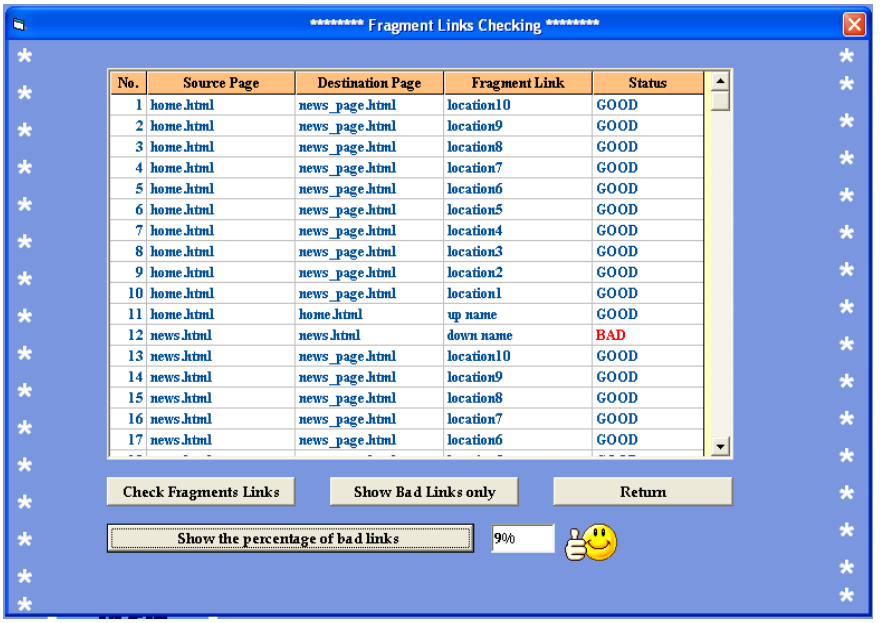

Figure 7
Relative links checking interface with information

5. Conclusion.

Broken links in a Web site is common problem on the Internet today. They inconvenience visitors, inhibit proper navigation of a site, prohibit access to Web site content and reduce the productivity of Web professionals.

In this paper a focused approach has been made to identify all possible broken links in the web site. This is an attempt to verify the quality assessment of Web Design.

In maintenance process, detected a broken link is needed to ensure that hyperlinks remain operational next week, next month, and next year.

Detecting a broken link system provides cost of time by repairing broken links.

We can further extend this work to identify other components of web site design for quality assessment which would further enable to improve the design as a part of the ideology of total quality management which emphasizes the continuous improvement of Design aspect and promote excellence of Web design.

6. References.

{1} Virtualis Glossary, "Virtualis System", 2001, available at http://www.virtualis.com/guides-glossary.html

{2} David M.,"CGI programming with Tcl", Addison-Wesley, 2000

{3} Saba A., "Internet and Arabic Search Engines", M. Sc. Thesis, Al-Nahrain University, 2002

{4} Wendy Lehnert, The Web Wizard's Guide to HTML, Addison Wesley, 2002

{5} Jennifer Niederst, Web Design in a Nutshell A Desktop Quick Reference 1st Edition, Clairemarie Fisher O'Leary, 1999.

{6} http://www.ichnet.org/glossary.htm

{7} Ingham D., B., Canghey S. J. and Litlle M.C., "Fixing the Broken Link Problem", United Kingdom, 1995.

{8} } G. Sreedhar, Dr. A.A. Chari, Dr.V.V.Venkata Ramana, A Qualitative and Quantitative Frame Work for effective Website Design, International Journal of Computer Applications (0975 – 8887)Volume 2 – No.1, May 2010.

{9} The University of Newcastle, Australia, 20 June 2011, http://www.newcastle.edu.au/copyright.html

{10} Keith Sutherland, Understanding the Internet A Clear Guide to Internet Technologies, Butterworth-Heinemann, 2000.

{11} M.Asante and R.S. Sherratt, "Improvement of Link Failure Restoration Utilizing Multi Protocol Label Switching (MPLS) as a Means to Maintian Quality of Service (QOS)", Journal of Science and Technology, Kumasi, Ghana, Vol. 29, No. 2 August, 2009.

{12} Jeffery V., "The Art and Science of Web Design", 2nd Ed, United state of America, 2001.

الكشف عن الرابط المقطوع في موقع الويب

الدكتور جمال فاضل توفيق* الدكتورعبد المنعم صالح رحمة** احسان قحطان احمد*

الملخص

ان الظاهرة السائدة في الانترنت اليوم هي الصلات المكسورة في مواقع الويب. حيث هذه الظاهرة تزعج زوار الموقع وتمنعهم من الابحار فيه وكذلك تمنعهم من الوصول الى محتوياته وتخفض من انتاجية محترفي الويب. علاوة على ذلك فان مجتمعنا اصبح اكثر اعتمادية على الاترنت. لذا فان التأكد من دقة وكفاءة والصيانة الدورية للصلات تكون لها الاولية عند صيانة الموقع.

ان الهدف من هذا البحث هوايجاد اسلوب لفحص كافة صلات الموقع، وتعطينا تقرير مفصل عن كافة الصلات الكسورة ان وجدت، ان كانت لصلات نص او صورة.

* جامعة النهرين
** الجامعة التكنولوجية