Visualization of Link Structures and URL Retrievals Utilizing Internal Structure of URLs Based on Brunch and Bound Algorithms

Personal Information Collection Robot

Kohei Arai[1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

Abstract— **Method for visualization of URL link structure and URL retrievals using internal structure of URLs based on brunch and bound method is proposed. Twisting link structure of URLs can be solved by the proposed visualization method. Also some improvements are observed for the proposed brunch and bound based method in comparison to the conventional URL retrieval methods.**

Keywords- visualization of link structure; URL retrieval; brunch and bound method; serch engine; information collection robot.

I. INTRODUCTION

Search engines, Google, Yahoo!, Baidu, Bing, Yandex, Ask, and AOL (market share order) are widely available for URL retrievals [1]-[3]. Information collection robots are working for creation of URL database which allows for smart information retrievals, in particular, URL retrievals. The conventional search engines can be divided into four categories, Directory types, Robot types, Meta types, and Hybrid types. Yahoo search engine is one of Hybrid types (Robot type and manmade classification) while Google search engine is one of Robot types. Any of search engines classify newly created web sites into some classes in the database by human and/or robots. Thousands of hits are obtained when users try to search the most preferable web site with a single keyword or frequent keywords. On the other hands, not appropriate web sites are hit by using simple keywords. To make URL or information search much efficient in a comprehensive manner, the methods for visualization and for search with brunch and bound method are proposed. Furthermore, the proposed information collection robot is a personalized by users and does work together with any kinds of search engines.

In the database, all the URLs are linked each other, in general. Visualization tools for URL link structure are very useful for not only manual URL search but also direct understanding the relations among URLs. Natto View is one of those of visualization tools. It is sometime hard to understand the relations among URLs for the reason that those are linked each other in twisted structure. The proposed visualization tool allows represent the twisted structure of the relations among URLs.

By using the relations among URL links, it can be accelerate information retrievals, or URL retrievals based on brunch and bound method. Through experiments with internal URL links in Saga University, it is verified that the proposed URL retrieval method is superior to the conventional URL retrieval methods.

The following section describes the proposed URL link structure visualization tool and URL retrieval method based on brunch and bound method followed by some experiments. Then conclusion is described together with some discussions.

II. PROPOSED METHOD

A. *Search Engine*

Configuration of the conventional search engines with information collection robot are shown in Figure 1.

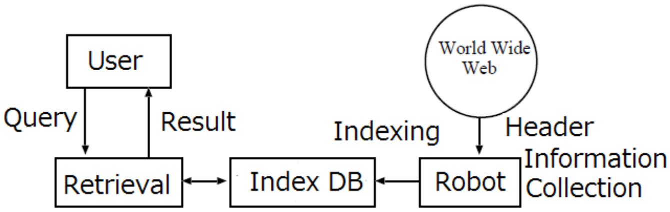


Figure 1 Configuration of the conventional search engines with information collection robot

Users send their query to the retrieval system. In parallel, information collection robot always collect header information from the web sites then create index for search in the index database (DB). Therefore, retrieval system send back the retrieved results by referencing the index DB.

B. *Process Flow*

Figure 2 shows procedure of the proposed URL search engine. The basic idea of the proposed search procedure is personalized information collection robot. Every time users search web site, information collection robot gathers preferable information of web sites and update users own personal database. Through learning processes, the personal information collection robot is getting craver. The other proposed method is visualization tool for a comprehensive representation of link relations among web sites. Then brunch and bound method based search is described.

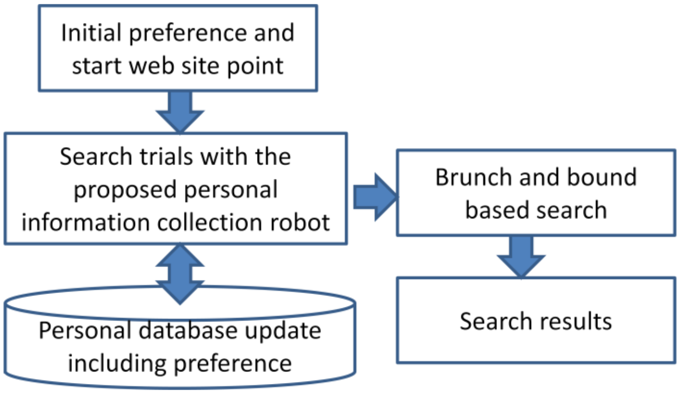

Initial preference and start web site point
Search trials with the proposed personal information collection robot
Brunch and bound based search
Personal database update including preference
Search results

Figure 2 Process flow of the proposed URL search engine

C. URL Link Structure Visualization Method

General information collecting robots start from the original URL, and get the header information which contains all the information required for links as well as keyword, descriptions and so on as tag information. Then the robot gathers the information required for links to the other URLs from the found URLs. Thus the gathered information required for reaching to the other related URLs are stored in the database of search engine.

There are some related research works on visualization of the relations among URLs, Kinematic model, Natto view model, and corn tree model. In the Kinematic model, there are the following conditions,

(1) The distance between spheres (nodes) has to be greater than the acceptable distance,

(2) Summation of the arcs (links) has to be minimized

(3) Avoid any overlap between arc and the different nearby nodes.

On the other hands, Natto view allows manipulate nodes and links manually. Magnification, translation, rotation are available to improve visibility. Meanwhile, corn tree model allows representation of hierarchical structure of the URLs using corn shaped three dimensional space. At the top of the corn, there is a parent URL followed by children and grandchildren and so on. Thus all the previously proposed visualization tools allow representation of focusing URLs in concern precisely.

A concept of the visualization of the relations among URLs is shown in Figure 3. All the URLs are illustrated with circles while the relations are represented with lines between URLs.

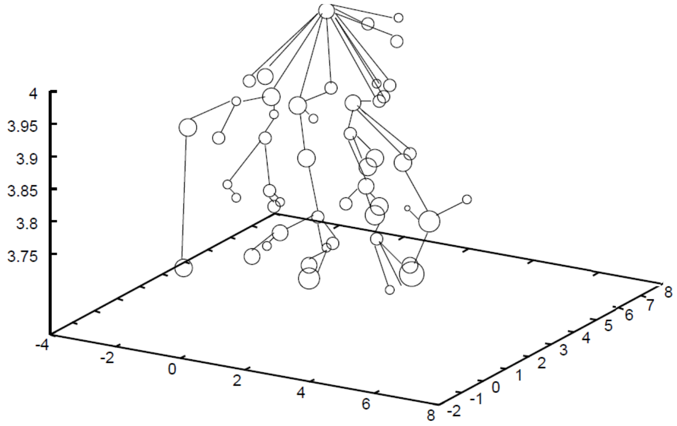


Figure 3 Concept of the visualization of the relations among URLs

In order to visualize the relations among URLs, the size and the distance are defined. The size defined as the number of links of the URL in concern. Deepening on the size, radius of node circle is determined. The distance between node circles is determined by the number of identical keywords and the words in the Meta tag of the header information of URL which must be highly correlated to the relation between URLs. The location of URL in X, Y plane is determined with the relations of web servers. If the URLs are in the same web server, then Z axis is determined while the URLs are in the different web server, then X,Y coordinates are determined as shown in Figure 4.

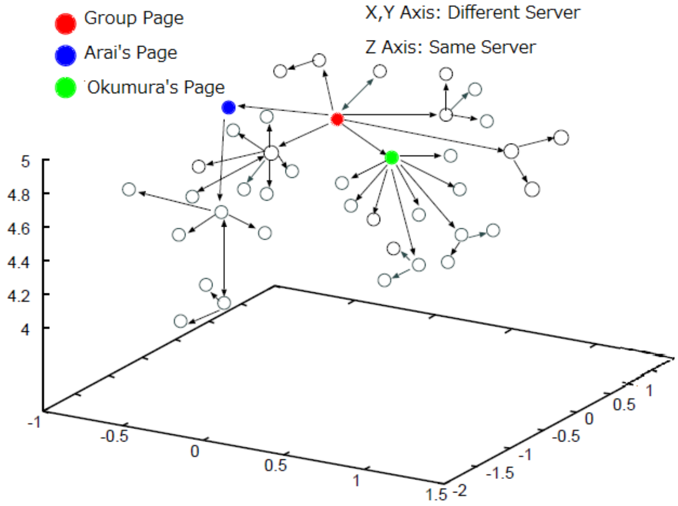


Figure 4 Determination of the location of URL node circles.

Then search begins from the red circle as shown in Figure 5, for instance. Information collection robot works next in the same time at which search engine is working.

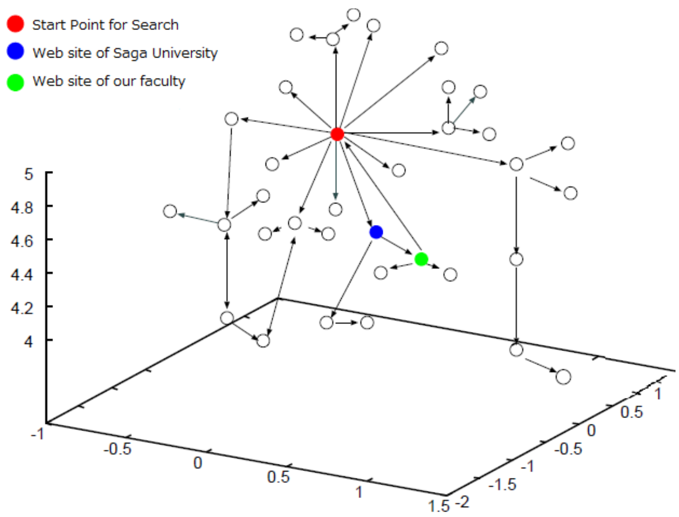
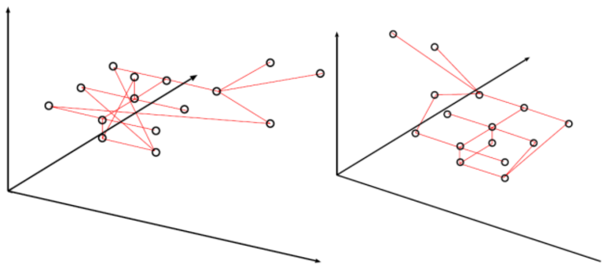


Figure 5 Information collection robot gather the information required for link and the next search

There is a problem on the twisted link representations as shown in Figure 6 (a). It is really hard to understand the links for the twisted links. By using Z axis allowance. untwisted representation can be done as shown in Figure 6 (b)



(a)Twisted links (b)Untwisted links

Figure 6 Method for untwisted representation of links among URL nodes

D. Information Collection Robot

The proposed information collection robot gathers header information together with checking whether or not information contains users' preferable information. Keywords and the frequency of the keywords are defined as users' preferences. Therefore, users can reach the most preferable web site efficiently by keying these preferences. Keywords and their frequencies are linked together. Therefore, users can omit key-in the frequency. Also information collection robot has learning capability. Therefore, only thing users have to do is just key-in their preferable keywords.

A Client-Server system is assumed for information collection robot. With the socket interface, server and client send and receive information. Hand shake procedure is shown below,

```
new IO::Socket::INET(
     [LocalAddr => 'hostname',]
     LocalPort => 'port',
     Proto => 'protocol',
     Listen => listen-limit,
     Reuse => reuse-number )
```

This is for the server side and the following is the client side,

```
new IO::Socket::INET (
     PeerAddr => 'hostname',
     PeerPort => 'port',
     Proto => 'protocol',
     TimeOut => 'timeout-second')
```

E. Brunch and Bound Search Method

URL search has to be done in an efficient manner. Blue force type of search method (try to search all available URLs) is not efficient. There is brunch bound search method as one of optimization methods which allows bound the brunch at which no appropriate brunch exists further below as shown in Figure 7.

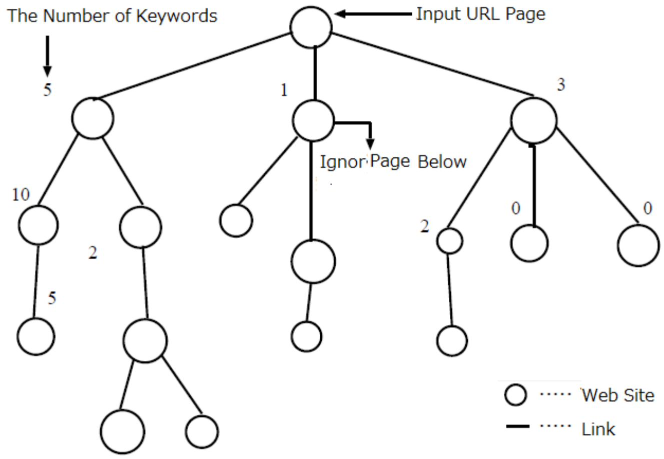


Figure 7 Brunch and bound search method

Namely, the procedure of the proposed brunch and bound is as follows,

(1) Client connect to URLs which are included in the designated web site through socket interface and then Web page contents are saved in the DB

(2) Check the frequency of the corresponding keywords from the saved web page contents in the DB

(3) If the frequency is not exceed the certain number, then the URL and URLs which are existing below the URL are omitted for search

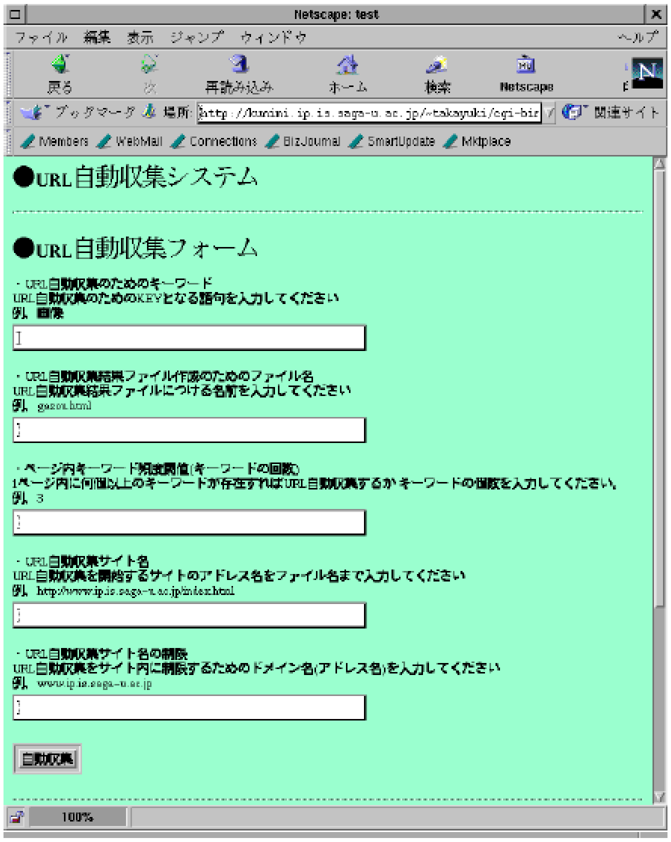
(4) (1) to (3) are performed for new query of search

(5) Retrieved results are output when the search is finished then the result file is converted to html file.

III. Implementation And Experiments

A. Web Design

Web pages are created with Perl programming language. Perl is interpreter language which allows input text and output report as well as text files manipulations and system management. Figure 8 shows an example of web page which allows input keywords for URL collection. Only thing users have to do is just key-in their own preferable keywords in the boxes.


Netscape: test
ファイル 編集 表示 ジャンプ ウィンドウ ヘルプ
戻る 次 再読み込み ホーム 検索 Netscape
ブックマーク 場所: http://kunimi.ip.is.saga-u.ac.jp/~takayuki/cgi-bir 関連サイト
Members WebMail Connections BizJournal SmartUpdate Mktplace
100%

●URL自動収集システム

●URL自動収集フォーム

・URL自動収集のためのキーワード
URL自動収集のためのKEYとなる語句を入力してください
例、画像

・URL自動収集結果ファイル作成のためのファイル名
URL自動収集結果ファイルにつける名前を入力してください
例、gazou.html

・ページ内キーワード頻度閾値(キーワードの回数)
1ページ内に何個以上のキーワードが存在すればURL自動収集するかキーワードの個数を入力してください。
例、3

・URL自動収集サイト名
URL自動収集を開始するサイトのアドレス名をファイル名まで入力してください
例、http://www.ip.is.saga-u.ac.jp/index.html

・URL自動収集サイト名の制限
URL自動収集をサイト内に制限するためのドメイン名(アドレス名)を入力してください
例、www.ip.is.saga-u.ac.jp

自動収集

Figure 8 Example of web page which allows input keywords for URL collection.

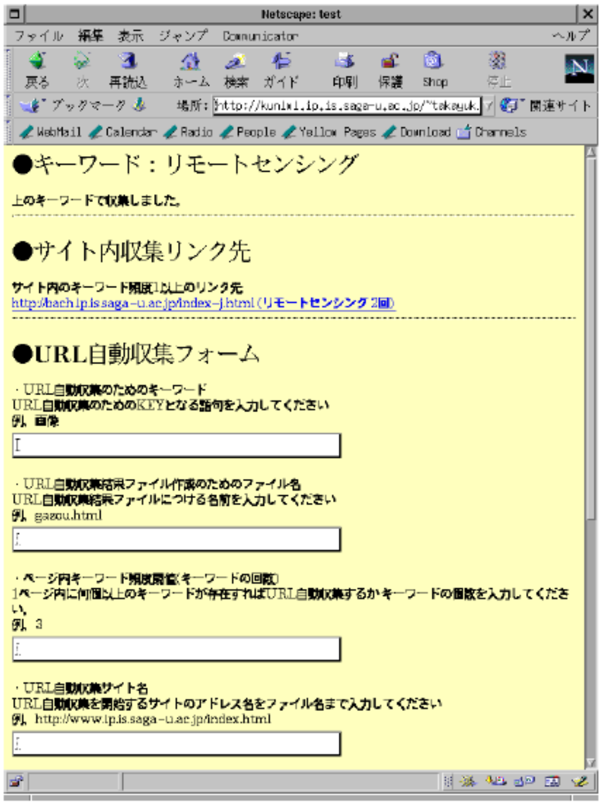
B. *Example of Search Results*

Figure 9 and 10 shows the retrieved results with the keywords, "Remote sensing" and "Image", respectively. All the existing URLs in the laboratory are listed as the results. For the search results of "Remote sensing", there are eight URLs as the search results while that for the keyword of "Image", there are 24 of URLs as the search results. The order of the search results of URLs are sorted by the frequency.

If users start their search with just "Remote Sensing" then users received Figure 9 (a) of search results. Then users selected the first URL, after that, users received Figure 9 (b) of search results. These are same thing for the keyword "Image". Thus users can search deeper and deeper as much as they could.

C. *Effect of Brunch and Bound Search Method*

In order to confirm the effect of the brunch and bound search method, man-machine time, CPU time, and elapsed time are evaluated for the search with the keywords "Remote Sensing" and "Image". The results are summarized in the Table1. It is quite obvious that the proposed brunch and bound search method is superior to the blue force type of search which search all the available URLs.


Netscape: test
ファイル 編集 表示 ジャンプ Communicator ヘルプ
戻る 次 再読込 ホーム 検索 ガイド 印刷 保護 Shop 停止
ブックマーク 場所: http://kunimi.ip.is.saga-u.ac.jp/~takayuk. 関連サイト
WebMail Calendar Radio People Yellow Pages Download Channels

●キーワード：リモートセンシング

上のキーワードで収集しました。

●サイト内収集リンク先

サイト内のキーワード頻度1以上のリンク先
http://bach.ip.is.saga-u.ac.jp/index-j.html (リモートセンシング 2回)

●URL自動収集フォーム

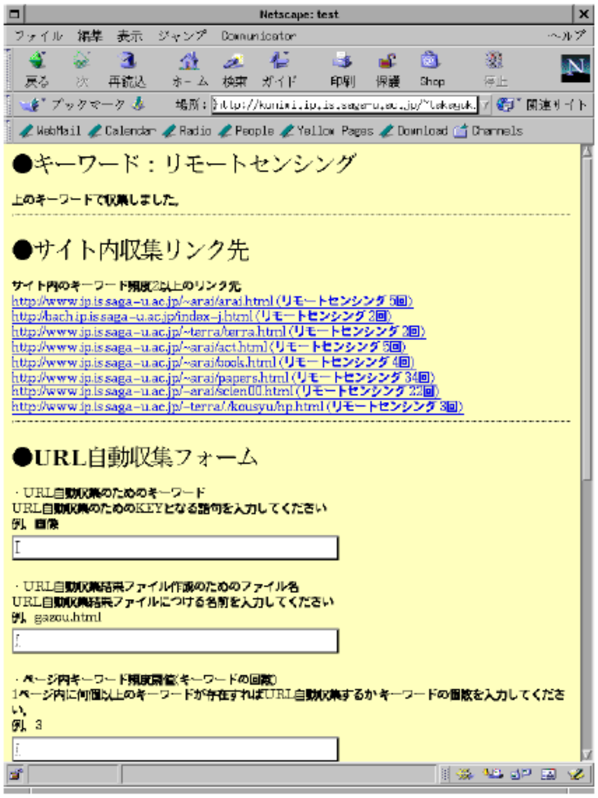
・URL自動収集のためのキーワード
URL自動収集のためのKEYとなる語句を入力してください
例、画像

・URL自動収集結果ファイル作成のためのファイル名
URL自動収集結果ファイルにつける名前を入力してください
例、gazou.html

・ページ内キーワード頻度閾値(キーワードの回数)
1ページ内に何個以上のキーワードが存在すればURL自動収集するかキーワードの個数を入力してください。
例、3

・URL自動収集サイト名
URL自動収集を開始するサイトのアドレス名をファイル名まで入力してください
例、http://www.ip.is.saga-u.ac.jp/index.html

(a)


Netscape: test
ファイル 編集 表示 ジャンプ Communicator ヘルプ
戻る 次 再読込 ホーム 検索 ガイド 印刷 保護 Shop 停止
ブックマーク 場所: http://kunimi.ip.is.saga-u.ac.jp/~takayuk. 関連サイト
WebMail Calendar Radio People Yellow Pages Download Channels

●キーワード：リモートセンシング

上のキーワードで収集しました。

●サイト内収集リンク先

サイト内のキーワード頻度2以上のリンク先
http://www.ip.is.saga-u.ac.jp/~arai/arai.html (リモートセンシング 5回)
http://bach.ip.is.saga-u.ac.jp/index-j.html (リモートセンシング 2回)
http://www.ip.is.saga-u.ac.jp/~terra/terra.html (リモートセンシング 2回)
http://www.ip.is.saga-u.ac.jp/~arai/act.html (リモートセンシング 5回)
http://www.ip.is.saga-u.ac.jp/~arai/book.html (リモートセンシング 4回)
http://www.ip.is.saga-u.ac.jp/~arai/papers.html (リモートセンシング 34回)
http://www.ip.is.saga-u.ac.jp/~arai/scien00.html (リモートセンシング 22回)
http://www.ip.is.saga-u.ac.jp/~terra/./kousyu/hp.html (リモートセンシング 3回)

●URL自動収集フォーム

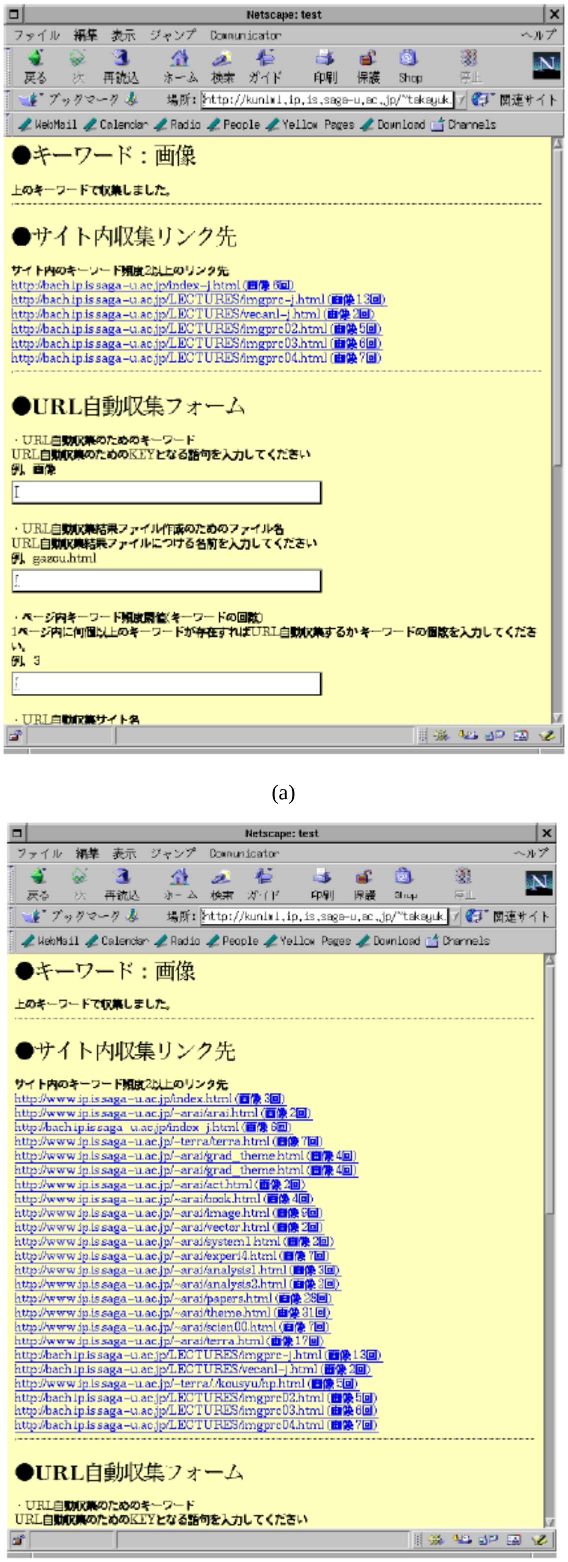
・URL自動収集のためのキーワード
URL自動収集のためのKEYとなる語句を入力してください
例、画像

・URL自動収集結果ファイル作成のためのファイル名
URL自動収集結果ファイルにつける名前を入力してください
例、gazou.html

・ページ内キーワード頻度閾値(キーワードの回数)
1ページ内に何個以上のキーワードが存在すればURL自動収集するかキーワードの個数を入力してください。
例、3

(b)

Figure 9 Retrieved results with the keyword "Remote Sensing"


Netscape: test
ファイル 編集 表示 ジャンプ Communicator ヘルプ
戻る 次 再読込 ホーム 検索 ガイド 印刷 保護 Shop 停止
ブックマーク 場所: http://kunimi.ip.is.saga-u.ac.jp/~takayuk 関連サイト
WebMail Calendar Radio People Yellow Pages Download Channels

●キーワード：画像

上のキーワードで収集しました。

●サイト内収集リンク先

サイト内のキーワード頻度2以上のリンク先
http://bach.ip.is.saga-u.ac.jp/index-j.html (画像 6回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc-j.html (画像 13回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/vecanl-j.html (画像 2回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc02.html (画像 5回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc03.html (画像 6回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc04.html (画像 7回)

●URL自動収集フォーム

・URL自動収集のためのキーワード
URL自動収集のためのKEYとなる語句を入力してください
例, 画像

・URL自動収集結果ファイル作成のためのファイル名
URL自動収集結果ファイルにつける名前を入力してください
例, gazou.html

・ページ内キーワード頻度閾値(キーワードの回数)
1ページ内に何個以上のキーワードが存在すればURL自動収集するかキーワードの個数を入力してください。
例, 3

・URL自動収集サイト名

(a)

Netscape: test
ファイル 編集 表示 ジャンプ Communicator ヘルプ
戻る 次 再読込 ホーム 検索 ガイド 印刷 保護 Shop 停止
ブックマーク 場所: http://kunimi.ip.is.saga-u.ac.jp/~takayuk 関連サイト
WebMail Calendar Radio People Yellow Pages Download Channels

●キーワード：画像

上のキーワードで収集しました。

●サイト内収集リンク先

サイト内のキーワード頻度2以上のリンク先
http://www.ip.is.saga-u.ac.jp/index.html (画像 3回)
http://www.ip.is.saga-u.ac.jp/~arai/arai.html (画像 2回)
http://bach.ip.is.saga u.ac.jp/index j.html (画像 6回)
http://www.ip.is.saga-u.ac.jp/~terra/terra.html (画像 7回)
http://www.ip.is.saga-u.ac.jp/~arai/grad_theme.html (画像 4回)
http://www.ip.is.saga-u.ac.jp/~arai/grad_theme.html (画像 4回)
http://www.ip.is.saga-u.ac.jp/~arai/act.html (画像 2回)
http://www.ip.is.saga-u.ac.jp/~arai/book.html (画像 4回)
http://www.ip.is.saga-u.ac.jp/~arai/image.html (画像 9回)
http://www.ip.is.saga-u.ac.jp/~arai/vector.html (画像 2回)
http://www.ip.is.saga-u.ac.jp/~arai/system1.html (画像 2回)
http://www.ip.is.saga-u.ac.jp/~arai/experi4.html (画像 7回)
http://www.ip.is.saga-u.ac.jp/~arai/analysis1.html (画像 3回)
http://www.ip.is.saga-u.ac.jp/~arai/analysis2.html (画像 3回)
http://www.ip.is.saga-u.ac.jp/~arai/papers.html (画像 28回)
http://www.ip.is.saga-u.ac.jp/~arai/theme.html (画像 31回)
http://www.ip.is.saga-u.ac.jp/~arai/scien00.html (画像 7回)
http://www.ip.is.saga-u.ac.jp/~arai/terra.html (画像 17回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc-j.html (画像 13回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/vecanl-j.html (画像 2回)
http://www.ip.is.saga-u.ac.jp/~terra/./kousyu/hp.html (画像 5回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc02.html (画像 5回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc03.html (画像 6回)
http://bach.ip.is.saga-u.ac.jp/LECTURES/imgprc04.html (画像 7回)

●URL自動収集フォーム

・URL自動収集のためのキーワード
URL自動収集のためのKEYとなる語句を入力してください

(b)

Figure 10 Retrieved results with the keyword “Image”

D. Effect of the Starting URL Point

The effect of the search starting URL points is also evaluated with the keywords “Remote Sensing” and “Image”. The results are shown in Table 2. The results show that efficient search can be done by starting the search with the closest URL to the desired URL.

TABLE I. TABLE TYPE STYLES

	Brunch and Bound	All URL Search
Man-Machine Time(min)	0.58	1.74
CPU Time (sec)	0.41	1.34
Elapsed Time (min)	1.07	3.32

TABLE II. TABLE TYPE STYLES

	Number of Keywords	
Keyword	Image	Remote Sensing
From Top of Laboratory	18	9
From Okumura's Page	2	1

It is easy to say that. It is not so easy to search. By looking at the link structure of the URLs using the proposed visualization method, users can find the starting URL point for the designated search purposes easily.

IV. CONCLUSION

Method for visualization of URL link structure and URL retrievals using internal structure of URLs based on brunch and bound method are proposed. Twisting link structure of URLs can be solved by the proposed visualization method. Also some improvements are observed for the proposed brunch and bound based method in comparison to the conventional URL retrieval methods

ACKNOWLEDGMENT

The author would like to thank Mr. Takayuki Sera for his effort to conduct experiments.

REFERENCES

[1] Pavliva, Halia (2012-04-02). "Yandex Internet Search Share Gains, Google Steady: Liveinternet". Bloomberg.com. http://www.bloomberg.com/news/2012-04-02/yandex-internet-search-share-gains-google-steady-liveinternet.html. Retrieved 2012-05-14.

[2] Segev, Elad (2010). Google and the Digital Divide: The Biases of Online Knowledge, Oxford: Chandos Publishing.

[3] Vaughan, L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes, Information Processing & Management, 40(4), 693-707.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science, and Technology of the University of Tokyo from 1974 to 1978 also was with National Space Development Agency of Japan (current JAXA) from 1979 to 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He was appointed professor at Department of Information Science, Saga University in 1990. He was appointed councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was also appointed councilor of Saga University from 2002 and 2003 followed by an executive councilor of the Remote Sensing Society of Japan for 2003 to 2005. He is an adjunct professor of University of Arizona, USA since 1998. He also was appointed vice chairman of the Commission “A” of ICSU/COSPAR in 2008. He wrote 30 books and published 332 journal papers.