Visualization of WWW Construction Using of Web Mining

Koei Shiratori, Tsuyoshi Otake and Mamoru Tanaka

Sophia University
7-1 kioi-cho, chiyoda-ku, Tokyo, Japan
Phone:+81-3-3238-3878, Fax:+81-3-3238-3321
Email: koei@mamoru.ee.sophia.ac.jp

Abstract— This work is about Web Mining that use large amount of data of WWW information. In this paper, we show visualize WWW site construction. To explain information visually is more effective to understand the feature, tendency of the information than to explain with text. So we develop an application that help user to understand Web construction and to be able to see Web correlations.

1. Introduction

There are much data that are worth to analyze as Data Mining, for Example the customer information that stored by POS one by one, the customer special quality that used risk analyze by finance. Data you can get free are various, in that data, there are Web information of WWW. Now Web page exist more than one hundred million and the scale of data quantity are correctly No1. This data of Web are treasure mine of information and to use this information we solve every problem. When we deal with this enormous Web data, we use Web search system. That system is text search. If you input a key word to that system, the Web site response itemize Web site that accord to key word. Now there are only itemize text search system. But the approaches of Web data is not only this. For Example Browser show only the extracted data you want to know.

To express the information visually is more effective than using some sentences. In general, WWW search engine shows results in text form. So it is difficult to grasp the information of the WWW sites, and we do not know whether there are any information we want to know if the results are sorted by importance that each search sites decide. This work helps us to grasp WWW structure visually.

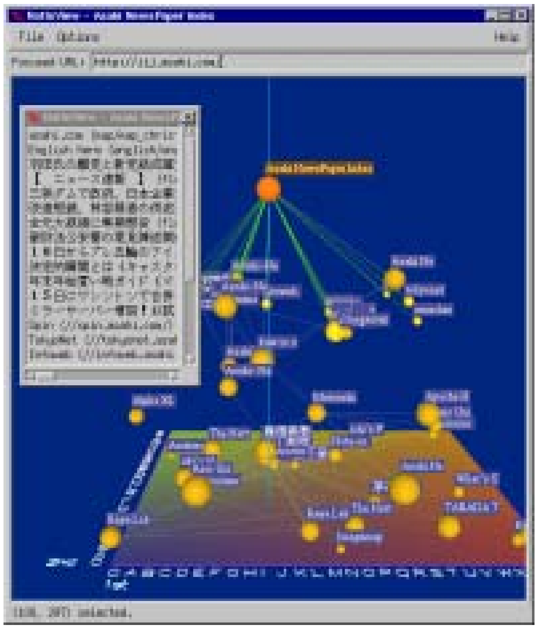

Figure 1: Natto View

2. Visualize Web space

Research of a visualization system is done from 1994. The visualization system of the link which using the 'Hyperbolic Tree' 'Web OOGL' is developed at Geometry Center in Minnesota University and published at VRML95[1]. The Nattoview is shown in fig.1. The ball and line which tie balls for the node of 3-dimensional space and the link between pages respectively expressing the WWW page. As the user lifts a focused node up, the nodes to which it links are lifted up together, and thus complicated networks are disentangled dynamically. By three dimensional perspective technique, the user can view both details of information connections near the selected node and global context of the large information space like fisheye lens model[2]. Lamping in Xerox PARC published the method is corresponding to change of a viewpoint using technique of 'Focus+Context'. This is commercialized as 'inXight'[3]. It is shown in fig.2. In Roma University 'Ptolomaeus' was published[4]. It is

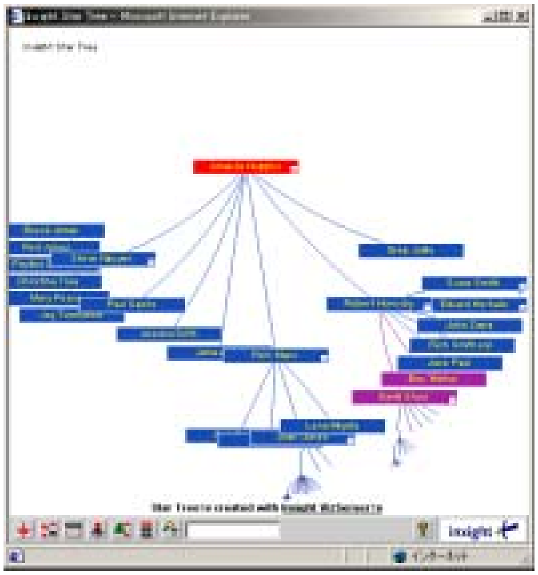

Figure 2: InXight

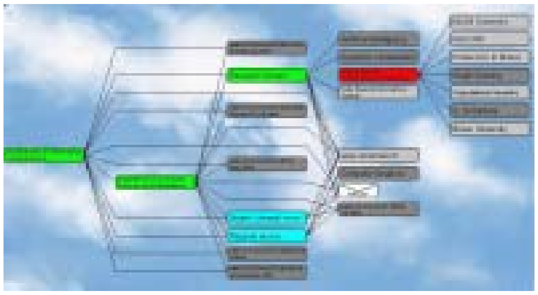

Figure 3: Ptolomaeus

shown in fig.3. The objects for these WWW visualization software are used by the home page of a company or a university. Hence, it can visualize the link information at one site. The WWW site of company is reconstituted by the visualization software after analyzing the structure of its HP.

3. The Arrangement of WWW sites to 3D space

To arrange the WWW sites on the 3D space has some advantages as follows.

- It is easy to grasp the construction of WWW sites.
- To use 3D space is able to grasp WWW sites more than to use 2D space.
- It shows the links of each WWW sites.

The indication of WWW data using 3D space explains at Figure.4.

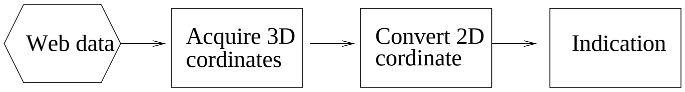

Figure 4: Indication flow

4. Module of WWW access engine

The module of WWW access engine acquires WWW files to trace hyper link. This module is constructed by 4 threads. The module is shown in Figure.5.

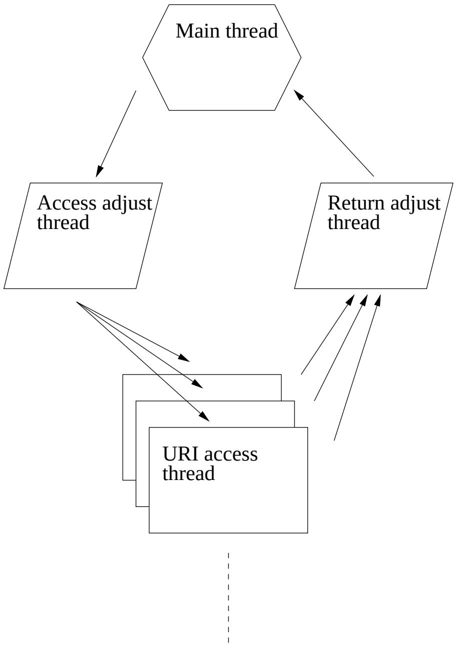

Figure 5: Relation of multi thread

The function of the 4 threads are:

- Main thread

  The application generates the main thread automatically, the one process requires at least one main thread. The main thread spawns the access adjust thread, the return adjust thread and the user interface. Also the main thread anayze the data.

- Access adjust thread

  This thread connects between the main thread and the URI access thread. The access adjust thread creates some URI access threads, the access adjust thread manages network access less than 3 URI access threads concurrently, because of considering network load.

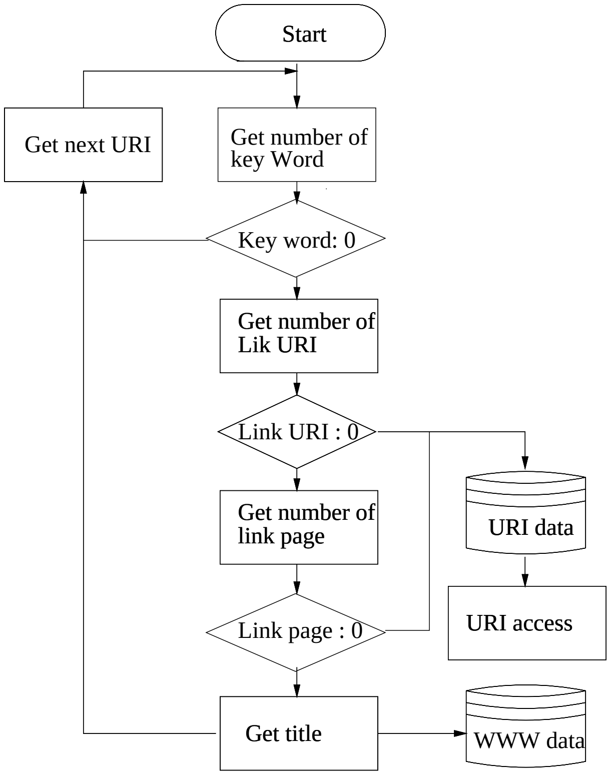

Figure 6: Extract Web information

- URI access thread

  This thread acquires WWW files from the network. This thread controls only one URI. The URI access thread is extinguished by the Access adjust thread after acquiring the data and giving it to The retun adjust thread.

- Return adjust thread

  Return adjust thread receives the data from URI access thread, then this thread transfers the data to the Main thread.

5. Extract information from WWW data

It is shown the flowchart of extracting information from WWW data in Figure.6.

- Acquisition of number of keyword

  The user inputs keyword, the Main thread investigates how many words match the keywords from each WWW file. If no word matches keyword, the Main thread do nothing, it changes into idle status until data has been derived from the Return adjust thread.

- Acquisition of page that hyper link is concentrated

  In many case, personal home page has the link page which hyper link collection of other WWW sites. The Main thread tries to find the link page from the acquired data. The thread search the characters of ”link” and ”Link” from html tag, it gets html body from them.

- Acquisition of hyper link

  The Main thread extracts the hyper link URI from WWW files.

6. Arrangement of correlation between WWW sites

In order to grasp the WWW sites easily, three kinds of WWW information are associated with three axes for display on three dimensions.

1. The number of link stretched

   In fact, many people visited the WWW sites, which have much useful information. Also such WWW sites are stretched by many hyper links. So the number of hyper links can be defined as one of the variable for evaluation.

2. The number of the internal hyper link on the top page

   The internal hyper link means the hyper link is stretched to WWW page at its WWW site. The large scale WWW site is tend to equip many internal hyper links on the top page. The number of internal hyper link can be defined as one of convenient index.

3. The number of keyword hit

   Both of the small scale WWW site and WWW site of stretched a few hyper links are categorized low ranking by the search sites, but sometimes such WWW site is publishing very useful information. So keyword matching is defined as one of variable for does not overlook precious WWW sites.

7. Simulation

First, we performed simulations of WWW sites visualization. 20 WWW sites are arranged on 3D space randomly, it is shown in Figure.7. The line between balls indicate hyper link.

Next, The number of the Z axis set to number of the stretched links, the radius of balls define number of the matched keyword. The simulation result is shown in Figure.8.

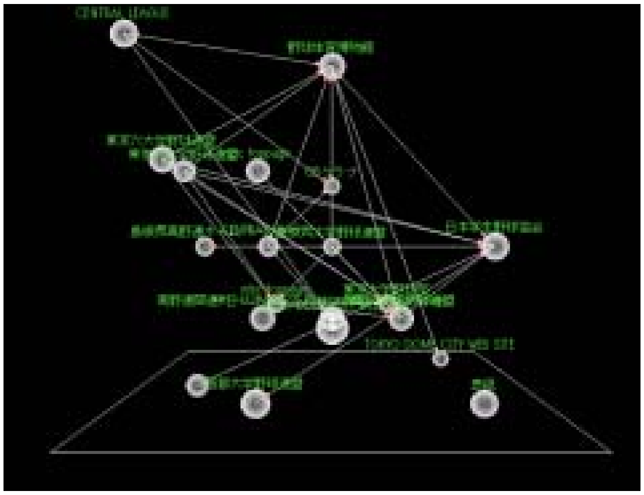

Figure 7: Visualization of WWW sites (Each WWW are arranged randomly)

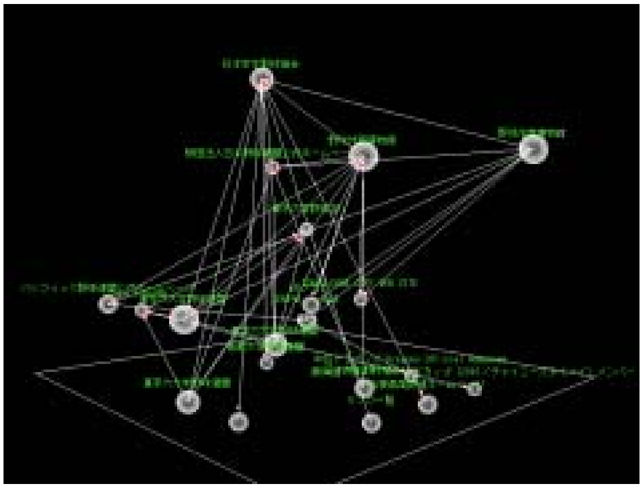

Figure 8: Visualization of WWW sites (Z axis relate with linked numbers)

To compare with Figure.8 to Figure.7, WWW sites are sorted into stretched link numbers on Figure.8, it is more comprehensive than Figure.7.

Finally, x, y, z axes assign the parameters as three information where it is defined on previous chapter, x is the the number of link stretched, y is the number of the internal hyper link on the top page, z is the number of keyword hit. The simulation result is shown in Figure.9. It seems that WWW sites are classified at several groups on Figure.9. The third simulation result most comprehensive other than simulation results.

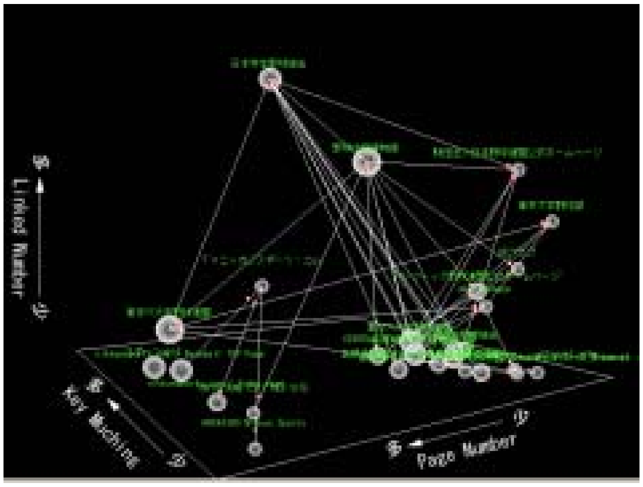

Figure 9: Visualization of WWW sites (Three axis relate with three information)

8. Conclusion

In this paper, we proposed to comprehend the characteristic of WWW sites which used by new application for visualizing WWW sites. It is realized by reconstruction using of Web mining architecture. We showed WWW sites correlation more comprehensive on 3D space.

References

[1] Munzner, T., and Burchard, P., "Visualizeing the Structure of the World Wide Web in 3D Hyperbolic Space" in Proc. of VRML'95, Computer Graphics, 33-38, 1995.

[2] H.Shiozawa,H.Nishiyama and Y.Matsushita, "The Natto View: An Architecture for Interactive Information Visualization" in Proc. of IPSJ 1997, vol38, No11.

[3] Lamping,J., Rao, R., and Pirolli, P., " A Focus+Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies" in Proc. of the ACM SIGGHI Conference on Human Factors in Computing Systems, 1995.

[4] Di Battista, G., Lillo, R., and Vernacotola, " Ptolomaeus: The Web Cartographer" in Proc. of GD'98, Springer LNCS 1547, 1998.