Since the early in 1990’s when frequent commercial use of the Internet starts, research on websites has been actively conducted both by academic researcher and by website practitioners. Three broad categories of the research are Web content mining, Web structure mining, and Web usage mining. Apart from those, there are some research on coloring, placement technique of images, texts, and links on each page. In this paper, we focus on the difference between two cluster structures of website, one induced from link-based property and the other from term-based property. The link-based property is stable until a new link is added, but the term-based one varies depending on the items for searching. We propose an evaluation method for website by comparing the structures of clusters resulted from these properties. As the clustering method, here we adopt kernel \( k \)-means method, and compare partial clusters derived from term-based property depending on the given sequence of particular terms to definite partial clusters from link-based property. In order to distinguish them, we try to adopt spectral analysis.

I. INTRODUCTION

Since commercial use of the internet began early in 1990s, information dissemination centering on websites has been done in many fields. In about 30 years, various techniques for Internet-based environment is developed and put to practical use, which gives websites the leading role in worldwide information dissemination. However, these years are also the period when unnecessary and meaningless information is accumulated on the website. Like as a database system that is not well maintained, websites within an organization may still keep information that needs to be expired or deleted, and there may be inappropriate links to them or from them. It is because the network administrator of the organizational website and the person in charge of each department’s home-page management can not properly manage pages, and there is no system for looking over and improving the whole site.

Regarding website links and described content, active research on search engines has been conducted and put to practical use. For example, as a hierarchical network search engine, HyPersuit proposed by [1], is well known and pioneering software in which content-link hyper-text clustering is exploited to obtain effective search results. The content-link hypertext clustering organizes documents into clusters of related documents based on the hybrid function of hyperlink and terms related structures. The hyperlink structure is described as the value of link similarity function of each of two nodes, and the term structure is described as the value of term-based similarity function using term frequency, size of document, and the term attributes. Chaomei Chen published a paper for analyzing the structure of a large hypermedia information space based on three types of similarity measures such as hypertext linkage, content similarity, and usage patterns [2, 3]. He tried to describe the network or its sub-networks as a graphical image of their Pathfinder Networks models, [4], based on pairwise integrated similarity by applying the vector space model [5, 6]. The primary purpose of our research is to develop an analyzing system for website from various perspectives with which website managers or webpage designers could improve the whole website by removing or adding proper links and pages. Of course, there is a problem as to what kind of website is good. Although excluding unnecessary links or information might be one of good website require-
ments, it is not always true from the user’s point of view. We already proposed in [7, 8] website evaluation system by combining some of website indexes and user evaluation indexes by applying user’s perspective information quality measures, [9, 10, 11]. In that research, we tried to construct a formula describing some of information quality measures as a formula with link or term based indexes. As an illustrative example, some coefficients of linear regression model between index values from a few dozens of pages in each website and information quality index obtained from the result of the factor analysis of questionnaire survey responses. However, the data volume was insufficient for both explanatory variables and objective variables, and the obtained result was not acceptable for practical use.

In this paper, while the user’s perspective evaluation is not involved, we create an application software written in Java language for collecting data from real website and analyzing link and term related properties. The rest of the paper organized as follows. Some of link and term related indexes are introduced in the next section, and analyzing method implemented or will be implemented in our application program are explained in the following. Some experimental results are described, then discuss them. The last part is the conclusion and about the future works.

II. INDEXES ON WEBSITE

Here we refer to two types of indexes on website, one is those related to the link structure and the other is those related to the term structure.

A. Link Related Index

Among various link related indexes proposed so far, Compactness and Stratum represent site-wide characteristics, Hub Weight and Authority Weight as characteristics of each node, and Complete Hyper Link Similarity (CHLS) represents the similarity of two nodes in terms of characteristics of links.

For an explanation of these indexes above, here we give assumptions and notations. Let \( P_1, \ldots, P_N \) be a set of entire pages as nodes at the target site, and put \( a_{ij} = 1 \) if there is a direct link from node \( P_i \) to node \( P_j \), otherwise put \( a_{ij} = 0 \), then consider the adjacent matrix \( A = (a_{ij}) \). Applying an appropriate method to this \( A \), find \( c_{ij} \) the shortest distance from \( P_i \) to \( P_j \), i.e., the minimum number of edges that must be traversed to reach. In case of being no route to reach, set a predefined sufficiently large number \( K \) for the value of \( c_{ij} \).

1) Compactness and stratum: As two indicators of site-wide status, here we refer to Compactness \( (C_p) \) and Stratum \( (St) \). Compactness is close to 1 if the nodes are densely connected, and is close to 0 otherwise, and Stratum is close to 1 when the whole node has close serial connections and is close to 0 otherwise. Therefore, these are complementary indexes, but as shown in (1) and (2), it is difficult to understand the clear relationship from their defining expressions.

\[
C_p = \frac{K}{K-1} - \frac{\sum c_{ij}}{(N^2 - N)(K-1)} \quad (1)
\]

\[
St = \sum_{i,j} |OD_j-ID_i| \quad \text{LAP}
\]

where \( OD_i = \sum_j c_{ij}, ID_i = \sum_j c_{ji}, \)

\[
\text{LAP} = \begin{cases} \frac{N^3}{N^2-N} & \text{(if } N \text{ is even) } \\ \frac{N^3}{N^2-N} & \text{(if } N \text{ is odd) } \end{cases}
\]

2) Hub weight and authority weight: For the importance value of each node as a referring or referred one in a directed graph, Kleinberg’s hub and authority weights are very popular [12]. Each component of the principal eigenvectors (eigenvector for the greatest positive real eigenvalue) of the matrices \( AA^t \) and \( A^tA \) represents the Hub Weight and the Authority Weight of the corresponding node.

In a homepage site, the number of nodes is several hundred to several thousand, or it exceeds 10,000 in some cases. These calculations are not so easy, thus we have not implemented them in our application software yet.

3) CHLS: The index proposed by Weiss et al. expresses the similarity between given two nodes by the link relation with other nodes and can be obtained by taking the weighted average of the following three values, [1].

\[
S^{expd}_{ij} = \frac{1}{2c_{ix}+1} + \frac{1}{2c_{ji}}
\]

\[
S^{anc}_{ij} = \sum_{x \in A_{ij}} \frac{1}{2(c_{ix}+c_{ji})}
\]

\[
S^{desc}_{ij} = \sum_{x \in A_{ij}} \frac{1}{2(c_{ix}+c_{ji})}
\]

Here, \( A_{ij} \) is a set of nodes such that there is at least one path to both \( P_i \) and \( P_j \), \( D_{ij} \) is a set of nodes such that there is at least one path from both \( P_i \) and \( P_j \), and \( c_{zi} \) is the shortest distance from \( P_z \) to \( P_i \) not passing \( P_j \).

From these values and weights \( w_x, w_a, w_d \), the CHLS index \( S_{ij}^{link} \) of each node pair \( (P_i, P_j) \) is defined as follows.

\[
S_{ij}^{link} = w_xS^{expd}_{ij} + w_aS^{anc}_{ij} + w_dS^{desc}_{ij} \quad (4)
\]

4) SG scores and graph spectrum: Shoda et al[13] considered all the connected sub-graph and calculate the total weight of each of them, then proposed to visually evaluate
the similarity by graphing their frequency of appearance as the spectrum.

For a weighted non-directed graph $G$ and the set of weight \( \{w(P)\} \), they considered all the connected sub-graphs \( \{SG \in 2^{G} : SG\} \) is connected and calculate the total weight of all node in $SG$ as \( w(SG) = \sum_{P \in SG} w(P) \) for each $SG$. Then, the graph spectrum is defined as the vector with component values of numbers of $SG$s whose weight are corresponding to the index number. In the paper, the graph spectrum is used to calculate the structural similarity of clusters corresponding to the index number. In the paper, the graph spectrum is used to calculate the structural similarity of clusters and apply $k$-means method to find a good cluster decomposition.

Here we propose the in- and out-weight of connected sub-graphs in a directed graph. For a fixed weights \( (w_1, w_2) \) with \( w_1 < w_2 \), the in-weight of a subgraph of only two nodes \( \{P_1, P_2\} \) is calculated as the weighted value \( w_{P_1 \rightarrow P_2} = w_1 w(P_1) + w_2 w(P_2) \) if there is a direct path from \( P_1 \) to \( P_2 \). Then define the in-weight for $SG$ of two nodes, called twin in-weight, to be the average of \( w_{P_1 \rightarrow P_2} \) and \( w_{P_2 \rightarrow P_1} \). When $SG$ has more than two nodes, the in-weight can be defined as the average of all the twin in-weights for connected twin subsets. However, the calculation efforts increase exponentially proportion to the number of nodes in $SG$.

**B. Term Related Index**

In order to define term related indexes for a set of fixed query terms \( q = \{Q_1, ..., Q_M\} \), we need the occurrence of \( Q_j \) in a page \( P \) denote by \( X_{P,Q_j} = 0 \) or 1), the term frequency, the number of \( Q_j \)'s appear in a page \( P \), denoted by \( TF_{P,Q_j} \), and the maximum number of \( TF_{P,Q_j} \) through all terms \( Q_j(j = 1, ..., M) \) denoted by \( TF_{P,max} \).

1) Boolean and vector spread activation: For some fixed values \( k_1 \) and \( k_2 \) satisfying \( 0 < k_2 < k_1 \), the Boolean spread Activation $R_{P,q}$ is defined:

$$R_{P,q} = \sum_{j=1}^{M} I_{P,Q_j},$$

where

$$I_{P,Q_j} = \begin{cases} k_1 & \text{if } X_{P,Q_j} = 1 \\ k_2 & \text{if } X_{P,Q_j} = 1 \text{ for some } P' \neq P \\ 0 & \text{otherwise} \end{cases}$$

The vector spread activation is defined by

$$RV_{P,q} = S_{P,q} + \sum_{P' \neq P} \alpha_{P,P'} S_{P',q}$$

where the reduced version of TFxIDF \( (S_{P,q}) \) is defined as follows,

$$S_{P,q} = \sum_{j=1}^{M} \frac{w_{P,Q_j}^{TF} \times IDF_{Q_j}}{n_{\text{edges}}}.$$  

With \( w_{P,Q_j}^{TF} \) and the "Inverse document frequency" \( IDF_{Q_j} \),

$$w_{P,Q_j}^{TF} = \frac{1}{2} \left( 1 + \frac{TF_{P,Q_j}}{TF_{P,max}} \right)$$

$$IDF_{Q_j} = \log \left( \frac{N}{\sum_{P' \neq P} X_{P',Q_j}} \right)$$

2) Term similarity index: In order to measure a similarity of two pages \( P_1 \) and \( P_2 \) for a particular set of search terms, the following index is defined as the term-based similarity index,

$$S_{P_1,P_2}^{\text{terms}} = \frac{\sum_{i=1}^{M} w_{P_1,Q_i} w_{P_2,Q_i}}{\sqrt{\sum_{i=1}^{M} w_{P_1,Q_i}^2 \sum_{i=1}^{M} w_{P_2,Q_i}^2}}$$

where \( w_{P_j,Q_i} = w_{P_j,Q_i}^{TF} \) with

$$w_{P_j,Q_i}^{TF} = \begin{cases} 10 & \text{if } Q_j \text{ is in title of } P_j \\ 5 & \text{if } Q_j \text{ is in headers or keywords or address in } P_j \\ 1 & \text{otherwise} \end{cases}$$

**III. CALCULATING AND ANALYZING METHOD**

Although the matrix $A$ can be obtained by examining the direct link relationship between nodes, we must find the shortest distance $c_{ij}$ from the node $P_i$ to the node $P_j$ for carrying out the calculation of the values of $C_i$, $S_t$, and $S^{\text{link}}$. For clarifying a given website as a directed graph, clustering methods are usually applied. Two types of clustering methods using both link-and term-related similarities are implemented in our program, and some others are planned to implement.

**A. Shortest Pass Calculation**

The problem of finding the shortest distance from a given node to another node is called the Single Source Shortest Pass (SSSP) problem, and the Dijkstra method [14] is well known as an algorithm to this problem. The original method required $O(n^2)$ execution time, and now improved to $O(e + n\log(n))$ ($e$ is the number of edges), [15].

On the other hand, the problem of finding the shortest distance for all node pairs at once is called the problem of All Pairs Shortest Pass (APSP). The Floyd-Warshall method
The shortest path from that node to the matrix $C(x, y)$ is given for all nodes of a directed graph whose edge weights are all 1. Suppose that a shortest distance matrix $C = (c_{ij})$ is given for all nodes of a directed graph whose edge weights are all 1. For any node pair $(x, z)$ with $x, z \in G$, put $c_{xz} = \infty$, when there is no path from $x$ to $z$.

For any $x, y, z \in G$ with $c_{xz} < \infty$, let $c_{xy}^z$ be the shortest distance from $x$ to $z$ that does not pass through $y$. Then follows:

1) Proposition 1: Suppose that a shortest distance matrix $C = (c_{xz})$ for a directed graph $(G, E)$ with edge weights all 1. For any node pair $(x, z)$ with $x, z \in G$, put $c_{xz} = \infty$, when there is no path from $x$ to $z$.

(i) If $c_{xz} \neq c_{xy} + c_{yz}$, then $c_{xz}^y = c_{xz}$.

(ii) If $c_{xz} = c_{xy} + c_{yz}$, and any $y'$ different from $y$ satisfies $c_{xy}' \neq c_{yz}$, then $c_{xz} = \infty$.

(iii) When there exists $c_{xz} = c_{xy} + c_{yz}$, for $y'$ giving $m = \min\{c_{y'z} - c_{yz} + c_{y'xy} : c_{y'xy} = c_{xy}\}$,

- If $c_{y'z} > m$, then $c_{xz}^y = c_{xz} + m$.
- If $c_{y'z} = m$, then necessary and sufficient condition for $c_{xz}^y = c_{xz}$ + $m$ is that there is a $y''$ different from $y$ satisfying $c_{y''z} + m = m$ and $c_{y''z} = c_{yz}$.

(Proof).

If there is a shortest distance through $y$, then $y$ will be its relay node, so $c_{xz} = c_{xy} + c_{yz}$, so (i) is obvious. When $c_{xz} = c_{xy} + c_{yz}$, $y$ is a node on the path giving the shortest distance. Since the weight of each edge is 1 there exists a node $y'$ for which $c_{xy'} = c_{xy}$ on the shortest path not passing $y$, thus (ii) is also valid. When there is a node $y'$ such that $c_{xz} = c_{xy} + c_{yz}$ and $c_{xy'} = c_{xy}$, then $c_{y'z} = c_{yz}$. If $y$ is on the shortest path from $y'$ to $z$, then $c_{y'y} = m$ holds from the condition $m = c_{y'y} - c_{yz}$. Together with $c_{y'y} > m$, we have $c_{xz}^y = c_{xz} + m$. In case that $c_{y'y}$ = $m$, we can determine $c_{xz}^y$. $c_{xz} + m$ depending on whether there is a node or not different from $y$ on the route giving the shortest distance $c_{y'y}$ and distance from $y'$ is $m$.

B. Clustering Methods

The clustering is a very important concept for the data structure analysis as data mining, and there are many proposals for general purpose algorithms and for specific data. Depending on the difference in data structure that is output as a result, they are divided into hierarchical algorithm and non-hierarchical one. The former outputs dendrogram and the latter outputs cluster set.

When a data set is well-defined in data space, clustering is performed by the distance, but when it is difficult to define the distance like nodes on a website, the similarity between two nodes as seen in the previous section is used. Here we refer to two of non-hierarchical clustering algorithms, one is the kernel version of the popular k-means algorithm and the other is SCAN.

1) Kernel K-means and SCAN: We quote MacQueen’s commentary on k-means procedure in [19], “Stated informally, the k-means procedure consists of simply starting with $k$-groups each of which consists of a single random point, and thereafter adding each new point to the group whose mean the new point is nearest. After a point is added to a group, the mean of that group is adjusted in order to take account of the new point. Thus, at each stage the $k$-means are, in fact, the means of the groups they represent (hence the term k-means).”

Dhillon et al.,[20], consider the weighted kernel k-means clustering and show the connection between it and spectral clustering algorithm. Before describing a formula for the kernel k-means with feature map $\phi$ from a set of nodes $G$ to the Hilbert space $H$ of functions on $G$ over $R$, let us review the definition of a kernel.

A positive definite kernel is a map

$$k : G \times G \rightarrow R$$

satisfying

- $k(x, y) = k(y, x)$ for any $x, y \in G$
- For any $\{c_i\} \subset R$ and any $\{x_i\} \subset G$,

$$\sum_{i=1}^{k} c_i \cdot k(x_i, x_j) \geq 0$$

Now go back to the kernel k-means algorithm. For a given number of clusters $k$, we try to find a set of clusters $C_1, ..., C_k$ minimizing the following value.

$$\sum_{i=1}^{k} \sum_{x \in C_i} ||\phi(x) - \mu_i||^2,$$

(7)
Where

\[
\|\phi(x) - \mu_i\|^2 = \left\| \phi(x) - \frac{1}{|C_i|} \sum_{x' \in C_i} \phi(x') \right\|^2
\]

\[
= k(x, x) - \frac{2}{|C_i|} \sum_{x' \in C_i} k(x, x')
\]

\[
+ \frac{1}{|C_i|^2} \sum_{x'' \in C_i} \sum_{x' \in C_i} k(x'', x')
\]

Xu et al., [21], proposed SCAN (Structural Clustering Algorithm for Networks), in [21] which, outputs three types of clusters such as “hub”, “outlier”, and ordinal clusters, by using structural similarity and two parameters \(0 \leq \epsilon \leq 1\) and \(\mu \in \mathbb{Z}^+\).

The structural similarity of two nodes \(x\) and \(x'\) is a number between 0 and 1 with value 1 only when \(x = x'\). A node \(x\) is called a core with respect to \(\epsilon\) and \(\mu\), if \(x\) has at least \(\mu\) number of nodes to which the structural similarity is greater than \(\epsilon\). Two nodes \(x\) and \(x'\) are connected when there exists one core \(x''\) from which they can be reached by following cores. Then output clusters are maximum subset of nodes any two of which are connected with each other. The hub is a node not belonging to any cluster and there are at least two distinct clusters each of which has a node with edge to it. When a node not belonging to any cluster does not satisfy the condition for the hub, it is called an outlier.

**IV. OUTLINE OF OUR APPLICATION PROGRAM**

Now we describe an outline of the developing application implementing the index calculation and the analyzing methods explained in the previous sections. The initial window is shown in the Figure 1. below composed of three essential panes and the menu bar.

![Initial Window](image)

**A. Main Window**

First of all, we submit the target website in the text field on the top of the main window, then starts retrieving the page by clicking the button below left. All the inner linked pages are retrieved while displaying the status in the text area below right with sequentially numbered page names. The left pane is for the key words or particular type of file search, and the results are displayed in the text area.

**B. Menus**

We set four menu items so far such as “FILE”, “Operations”, “Index Calc(ulation)”, and “Cluster Analysis”, as shown in Figure 2. Our program creates two types of data set, one is a text type file of home page written in html or xml language, the other is JAVA’s class object type data listed below the Figure 2. The second type of data set may require a lot of memory, so clearing memories are sometime necessary when new object data is reloaded.
Class Objects defined in Initial Page
static Floyd_Warshall
static InnerWebPG inPG;
static ArrayList<InnerWebPG> innerPGs;
static HashSet<String> totalOutLinkS;
static Floyd_Warshall dPG;
static int[] bsa;
static float[] vsa;
double [][] sim_Link;
double [][] sim_Term;

In the “Operations” item, the last sub-item for calculation of a distance matrix from an adjacent matrix by Floyd-Warshall method is essential to proceed a following step. The item “Index Calc” includes sub-items for calculation of indexes explained in the section 2, and it is saving or retrieving resulted data files are put here. The method of Prop. 1 is applied when calculating the complete hyperlink similarities. Here noticing that some search words should be input in the left pane of the main pane when calculating Boolean or vector spread activation indexes or term similarities. The last menu item is “Cluster Analysis” including two clustering methods depending on types of similarity measures. When choosing “Hard K-Means” method using link similarity measure, new window appears with two regions in the main pane as shown in the Figure 3. Enter two numbers in left text fields, one is the number of clusters and the other is the number of iterations, then k-means clustering method is applied using similarity measures as a kernel value by clicking “Start Cluster Calculation” button. The resulted clusters are shown in the text area, and a graph displaying each node placed on the circumference and direct links represented by lines appears in the right side of the text area by clicking the button below. To know the link status for each node, input the node number in text fields just above the “→Lines from the Nodes” or “←Lines into the Nodes” button. When vector spread activation values are given with some search key words, a circle with a radius in order of the magnitude of the score value is displayed near each of the node number.
In the cluster window, we set three menu items so far such as “File”, “Analyze”, and “Graphic”, as shown in Figure 4. By clicking the cluster wise analysis sub-item, cluster number buttons appear in the left pane by which adjacent link relations in the chosen cluster are graphically displayed. By the cluster pair analysis, Welch test for each of possible pairs of clusters using vector spread activation score values with significance level 1% and 5% are performed. In the Graphic item, we only have spectral chart sub-item, by which in- or out- spectral charts of chosen pair of clusters are displayed instead of direct link image.

Fig. 4. Cluster menu list

V. EXPERIMENTAL RESULTS

As experimental applications of our developing program, several websites of Japanese College are examined in October 2018 with searching key words on recruitment of students Figure 5. represents some of indexes values for one of example website consisting of 82 webpages. The resulted values of compactness 0.926 and stratum 0.022 imply that pages in the website are densely connected with each other. Pairs of Boolean spread activation and vector spread activation are displayed under the stratum value with the page number from 0 to 82. Although pages with higher Boolean spread activation value tend to have higher vector spread activation value, they are not always in proportion. For instance, (PG0: 23, 0.375), (PG46: 33, 0.538), (PG56: 23, 0.772), and (PG76: 3, 0.392). The following data set is that of link similarities ranging from 0 to 1, which are used for clustering.

Fig. 5. Index calculation

In the $k$-means clustering window, setting 3 as the cluster number and 20 as the number of iterations, we have clusters shown in the figure 6 where colored bubbles represent clusters and their vector spread activation score values as the size of the radius. Moreover, multiple lines coming out of page number 63 to the right represent the existence of direct link from the page of number in the input space to the others, and two lines to the left down are opposite direction.
We can see that there are few direct links from pages in cluster 3, and this cluster is the set of pages with least vector activation scores values.

The graphical images of each cluster are shown in the Figure 7, Figure 8, and Figure 9, from which we can see that the link relation only in the cluster 2 has relatively symmetry. From the results of Welch test, shown in the center pane of Figure 8, the average values of vector spread activation scores are differ by at least 5% significance level.
For the spectral calculation, we implemented up to only triple type of spectral so far. Figure 10 and Figure 11 describe each of single, twin, and triple types of in- and out-spectral which are normalized to have value between 0 and 1 for the cluster 2 and for the cluster 3 respectively. And the Figure 12 shows the comparison of out-spectral between cluster 2 and cluster 3. The cosine similarity indexes of in- and out-spectral vectors for each of cluster are also shown in the middle pane.
VI. CONCLUSION

We have investigated several indexes on website and analyzing method originated in the graph theory and clustering. Then developed an application software implementing some of them for finding link and term related properties in order to lead to an improvement of the website. We also give an experimental example with some discussion. The application software has not completed yet, and we intend to implement some other functions such as fuzzy clustering, pathfinder networks, etc. We need gather much
more website information to compare with each other to see better website properties. Further consideration is needed on some indicators, especially the interpretations of in- and out- spectral might be a big problem. As we refer in the 1st section, evaluation from users are also necessary to see what kind of indexes are related to so-called a good website.

REFERENCES


