Term-based Website Evaluation Applying Word Vectors and Readability Measures

Kiyoshi Nagata
Faculty of Business Administration, Daito Bunka University, Tokyo, Japan

Keywords: Website Evaluation, Similarity Measure, Word Vector, Readability Measures.

Abstract: Now the homepage is an important means of transmitting information not only in companies but also in any type of organization. However, it cannot be said that the page structure in the website is always in an appropriate state. Research on websites has been actively conducted both from academic and practical aspects, and sometime from three major categories such as web content mining, web structure mining, and web usage mining. In this paper, we mainly focus on term-based properties and propose a system that evaluates the appropriateness of link relationships in a given site taking those content and link structure properties into consideration. We also consider readability of text in each webpage by applying some of readability measures to evaluate the uniformity of them across all pages. We implement those systems in our previously developed multilingual support application, and some results by applying it to several websites are shown.

1 INTRODUCTION

In about 30 years since when commercial use of the internet began early in 1990s, information dissemination centering on websites has been developed in many fields by using various techniques for Internet-based environment. In this way, a lot of information is now accumulated on the Internet and is used all over the world.

On the other hand, webpages in an organization are sometimes expanded in an ad hoc way with the result that there are many pages containing unnecessary or not updated information. Network administrators focus on the smooth functioning of networks within an organization and are often not responsible for overall integrity or consistency. Especially in a large organization, the contents to be described on the homepage will be created independently, and each department decides the setting of the link in the organization. Such a method is consistent with the initial concept of the Internet such as ”no administrator required”, and may be supported by the idea that a collection of miscellaneous information is the Internet. However, compared to general information dissemination such as in SNS, the purpose of websites operated by organizations should be clarified, and for that purpose, overall coherence and efficiency should be emphasized.

From this point of view, it is desirable that the entire structure is constructed in consideration of the link structure between webpages and of the content relations of documents. Although it is unrealistic to completely rebuild a website that currently has hundreds, thousands, and possibly tens of thousands of pages and links according to the overall policy, it is required to analyze the pages on the website in the organization by the link structure and the document content structure, and to find out the problem by grasping the whole. We already proposed a system incorporating several website evaluation indicators, including existing ones, and adapted it to actual websites, but term-based evaluations were not sufficient inadequate (Nagata, 2019).

In this paper, we mainly propose the clustering system improved by following three types of values. One is a method of finding a spectrum for an effective graph using Kleinberg’s hub and authority weights (Kleinberg, 1999), the second is a similarity evaluation method using an extended keyword search index using word vector representation (Mikolov et al., 2013; Mikolov et al., 2015), and the third is using the readability index of the document.

The rest of paper is organized as follows. The application program already constructed by implementing some of the links and the terms related indexes is introduced while showing the execution result and pointing out the problem in the next section. In the following sections, the improvement methods for the

Nagata, K.
Term-based Website Evaluation Applying Word Vectors and Readability Measures.
DOI: 10.5220/0011383900003318
In *Proceedings of the 18th International Conference on Web Information Systems and Technologies (WEBIST 2022)*, pages 241-248
ISBN: 978-989-758-613-2; ISSN: 2184-3252

problem will be described one by one. We also describe how to implement some of the texts readability indexes. The last part is the conclusion and future works.

2 EXISTING SYSTEM AND PROBLEMS

In this section, the outline of the already developed system is described, and the execution results of the application program and some problems will be pointed out.

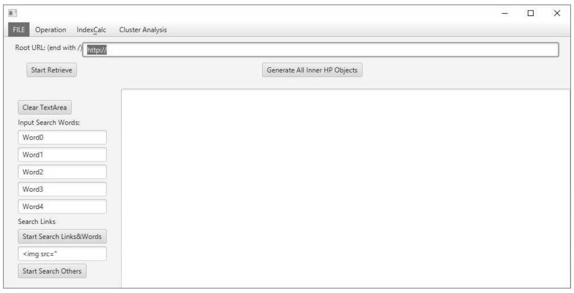
2.1 Overview of Website Evaluation and Our System

As a hierarchical network search engine, HyPersuit proposed by Ron Weiss et al. (Weiss et al., 1996) is well known and pioneering software. In the HyPersuit, the content-link hypertext clustering which organizes documents into clusters of related documents based on the hybrid function of hyperlink and terms related structures is exploited to obtain effective search results. To describe the hyperlink structure and the term structure, the value of link similarity functions of each of two nodes and term-based similarity function using term frequency, size of document, and the term attributes are applied respectively.

Chaomei Chen tried to describe the network or its sub-networks as a graphical image of the Pathfinder Networks models (Schvaneneldt et al., 1988) based on pairwise integrated similarity by applying the vector space model (Salton et al., 1975). He also 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 (Chen, 1998). Proposing website evaluation system by combining some of website indexes and user evaluation indexes by applying user's perspective information quality measures (Lee et al., 2002; Cherl and Locsin, 2018; Noorzad and Sato, 2017), we tried to construct a formula describing some of information quality measures as a formula with link- or term-based indexes (Liang and Nagata, 2011). In the paper (Nagata, 2019), an analyzing system for website from various perspectives id developed which helps website managers or designers improve the whole website by removing or adding proper links and pages. In the application program, while the user's perspective evaluation is not involved, collecting data from real website and analyzing link and term related properties are implemented (Nagata, 2019). The developing language is Java and the JavaFx library is used for the construction of graphical interface. Here we briefly describe some of core part of the previous system.

2.1.1 Retrieving Webpages

By giving a specific website URL, the system starts to retrieve almost all the webpages in side the website. Figure 1 shows the initial page of the system.



Root URL: (end with /) http://

Word0

Word1

Word2

Word3

Word4

<img src="

Figure 1: Initial Window.

For its implementation, we simply get the HTML or XML texts using the "URL" and "URLConnection" class objects, then analyzed the text to retrieve all the proper link tags. However, we needed to create the exception list of some types of link to be excluded, and the program sometimes failed retrieving at the stage of iterative processing especially in overseas websites.

2.1.2 Link-based Index

The basic indicators for the overall link relationship are Compactness (Cp) and Stratum (St), which must be complementary indexes. However as shown in (1), it is difficult to understand the clear relationship from their defining expressions.

$$\begin{cases} Cp &= \dfrac{K}{K-1} - \dfrac{\sum_{i,j} c_{ij}}{(N^2-N)(K-1)}, \\ St &= \dfrac{\sum_i |OD_i - ID_i|}{LAP}, \end{cases} \tag{1}$$

where K is a certain big number, c_{ij} represent the shortest distance from the page P_i to the page P_j and

$$OD_i = \sum_{j=1}^{N} c_{ij},\ ID_i = \sum_{j=1}^{N} c_{ji},$$

$$LAP = \begin{cases} \dfrac{N^3}{4} & (\text{if } N \text{ is even}), \\ \dfrac{N^3-N}{4} & (\text{if } N \text{ is odd}). \end{cases}$$

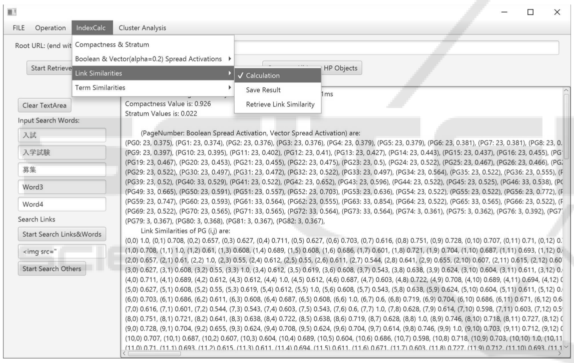
The Complete Hypertext Link Similarity($CHLS$) index is proposed by Weiss et al. to express the similarity between given two nodes by the link relation with other nodes, and *CHLS* index can be obtained by taking the weighted average of the following three values, (Weiss et al., 1996).

$$\begin{cases} S^{spl}_{ij} &= \dfrac{1}{2^{c_{ij}}} + \dfrac{1}{2^{c_{ji}}}, \\ S^{anc}_{ij} &= \displaystyle\sum_{x \in A_{ij}} \dfrac{1}{2^{(c^{\bar{j}}_{xi} + c^{\bar{i}}_{xj})}}, \\ S^{dsc}_{ij} &= \displaystyle\sum_{x \in D_{ij}} \dfrac{1}{2^{(c^{\bar{j}}_{ix} + c^{\bar{i}}_{jx})}}. \end{cases}$$

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 $c^{\bar{j}}_{xi}$ is the shortest distance from P_x to P_i not passing P_j. From these values and given weights w_s, w_a, w_d, the *CHLS* index S^{link}_{ij} of each node pair (P_i, P_j) is defined as follows.

$$S^{link}_{ij} = w_s S^{spl}_{ij} + w_a S^{anc}_{ij} + w_d S^{dsc}_{ij}. \tag{2}$$

Figure 2 shows the calculation results of the link similarities.



```
                                                1ms
Compactness Value is: 0.926
Stratum Values is: 0.022

    (PageNumber: Boolean Spread Activation, Vector Spread Activation) are:
(PG0: 23, 0.375), (PG1: 23, 0.374), (PG2: 23, 0.376), (PG3: 23, 0.376), (PG4: 23, 0.379), (PG5: 23, 0.379), (PG6: 23, 0.381), (PG7: 23, 0.381), (PG8: 23, 0.
(PG9: 23, 0.397), (PG10: 23, 0.395), (PG11: 23, 0.402), (PG12: 23, 0.41), (PG13: 23, 0.427), (PG14: 23, 0.443), (PG15: 23, 0.437), (PG16: 23, 0.455), (PG1
(PG19: 23, 0.467), (PG20: 23, 0.453), (PG21: 23, 0.455), (PG22: 23, 0.475), (PG23: 23, 0.5), (PG24: 23, 0.522), (PG25: 23, 0.467), (PG26: 23, 0.466), (PG2
(PG29: 23, 0.522), (PG30: 23, 0.497), (PG31: 23, 0.472), (PG32: 23, 0.522), (PG33: 23, 0.497), (PG34: 23, 0.564), (PG35: 23, 0.522), (PG36: 23, 0.555), (P
(PG39: 23, 0.52), (PG40: 33, 0.529), (PG41: 23, 0.522), (PG42: 23, 0.652), (PG43: 23, 0.596), (PG44: 23, 0.522), (PG45: 23, 0.525), (PG46: 33, 0.538), (PC
(PG49: 33, 0.665), (PG50: 23, 0.591), (PG51: 23, 0.557), (PG52: 23, 0.703), (PG53: 23, 0.636), (PG54: 23, 0.522), (PG55: 23, 0.522), (PG56: 23, 0.772), (P
(PG59: 23, 0.747), (PG60: 23, 0.593), (PG61: 33, 0.564), (PG62: 23, 0.555), (PG63: 33, 0.854), (PG64: 23, 0.522), (PG65: 33, 0.565), (PG66: 23, 0.522), (P
(PG69: 23, 0.522), (PG70: 23, 0.565), (PG71: 33, 0.565), (PG72: 33, 0.564), (PG73: 33, 0.564), (PG74: 3, 0.361), (PG75: 3, 0.362), (PG76: 3, 0.392), (PG7
(PG79: 3, 0.367), (PG80: 3, 0.368), (PG81: 3, 0.367), (PG82: 3, 0.367),
    Link Similarities of PG (i,j) are:
(0,0) 1.0, (0,1) 0.708, (0,2) 0.657, (0,3) 0.627, (0,4) 0.711, (0,5) 0.627, (0,6) 0.703, (0,7) 0.616, (0,8) 0.751, (0,9) 0.728, (0,10) 0.707, (0,11) 0.71, (0,12) 0.
(1,0) 0.708, (1,1) 1.0, (1,2) 0.61, (1,3) 0.608, (1,4) 0.689, (1,5) 0.608, (1,6) 0.686, (1,7) 0.601, (1,8) 0.721, (1,9) 0.704, (1,10) 0.687, (1,11) 0.693, (1,12) 0.
(2,0) 0.657, (2,1) 0.61, (2,2) 1.0, (2,3) 0.55, (2,4) 0.612, (2,5) 0.55, (2,6) 0.611, (2,7) 0.544, (2,8) 0.641, (2,9) 0.655, (2,10) 0.607, (2,11) 0.615, (2,12) 0.60
(3,0) 0.627, (3,1) 0.608, (3,2) 0.55, (3,3) 1.0, (3,4) 0.612, (3,5) 0.619, (3,6) 0.608, (3,7) 0.543, (3,8) 0.638, (3,9) 0.624, (3,10) 0.604, (3,11) 0.611, (3,12) 0.
(4,0) 0.711, (4,1) 0.689, (4,2) 0.612, (4,3) 0.612, (4,4) 1.0, (4,5) 0.612, (4,6) 0.687, (4,7) 0.603, (4,8) 0.722, (4,9) 0.708, (4,10) 0.689, (4,11) 0.694, (4,12) 0
(5,0) 0.627, (5,1) 0.608, (5,2) 0.55, (5,3) 0.619, (5,4) 0.612, (5,5) 1.0, (5,6) 0.608, (5,7) 0.543, (5,8) 0.638, (5,9) 0.624, (5,10) 0.604, (5,11) 0.611, (5,12) 0.
(6,0) 0.703, (6,1) 0.686, (6,2) 0.611, (6,3) 0.608, (6,4) 0.687, (6,5) 0.608, (6,6) 1.0, (6,7) 0.6, (6,8) 0.719, (6,9) 0.704, (6,10) 0.686, (6,11) 0.671, (6,12) 0.6
(7,0) 0.616, (7,1) 0.601, (7,2) 0.544, (7,3) 0.543, (7,4) 0.603, (7,5) 0.543, (7,6) 0.6, (7,7) 1.0, (7,8) 0.628, (7,9) 0.614, (7,10) 0.598, (7,11) 0.603, (7,12) 0.5
(8,0) 0.751, (8,1) 0.721, (8,2) 0.641, (8,3) 0.638, (8,4) 0.722, (8,5) 0.638, (8,6) 0.719, (8,7) 0.628, (8,8) 1.0, (8,9) 0.746, (8,10) 0.718, (8,11) 0.727, (8,12) 0
(9,0) 0.728, (9,1) 0.704, (9,2) 0.655, (9,3) 0.624, (9,4) 0.708, (9,5) 0.624, (9,6) 0.704, (9,7) 0.614, (9,8) 0.746, (9,9) 1.0, (9,10) 0.703, (9,11) 0.712, (9,12) 0
(10,0) 0.707, (10,1) 0.687, (10,2) 0.607, (10,3) 0.604, (10,4) 0.689, (10,5) 0.604, (10,6) 0.686, (10,7) 0.598, (10,8) 0.718, (10,9) 0.703, (10,10) 1.0, (10,11
(11,0) 0.71, (11,1) 0.693, (11,2) 0.615, (11,3) 0.611, (11,4) 0.694, (11,5) 0.611, (11,6) 0.671, (11,7) 0.603, (11,8) 0.727, (11,9) 0.712, (11,10) 0.693, (11,1
```

Figure 2: Link Similarity Calculation.

In the formula above, the shortest distance c_{ij} from page P_i to page P_j plays an important role. The shortest distance calculation problem is classified into two types such as Single Source Shortest Pass (*SSSP*) problem and All Pairs Shortest Pass (*APSP*) one. For the *SSSP* problem, the Dijkstra method (Dijkstra, 1959) is well known algorithm whose execution time improved to $O(e + n\log(\log n))$ (e is the number of edges), (Thorup, 2004). For the *APSP* problem, the Floyd-Warshall method (Floyd, 1962; Warshall, 1962) are known as an algorithm which requires $\Theta(n^3)$ complexity and a memory area proportional to n^2. We implemented this with the following procedure.

1. Let $C^{(0)}(= A)$ be a matrix representing (weighted) direct link relationships.
2. Node P_k $(k = 1, \ldots, N)$ is sequentially added as a transit node, and the (weighted) shortest distance changed for each node pair (P_i, P_j) is a component of the matrix $C^{(k)}$. When the shortest distance changes, P_k is stored as a predecessor node from P_i to P_j.

2.1.3 Term-based Index

As a term related nature of a page P, we implemented the vector spread activation defined by

$$RV_{P,q} = S_{P,q} + \sum_{P' \neq P} \alpha a_{PP'} S_{P',q}, \tag{3}$$

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

$$S_{P,q} = \sum_{j=1}^{M} w^{TF}_{P,Q_j} \times IDF_{Q_j}.$$

In the formula, we have w^{TF}_{P,Q_j} and the "inverse document frequency" IDF_{Q_j} defined by

$$\begin{cases} w^{TF}_{P,Q_j} = \dfrac{1}{2}\left(1 + \dfrac{TF_{P,Q_j}}{TF_{P,max}}\right) \\ IDF_{Q_j} = log\left(\dfrac{N}{\sum_{P'} X_{P',Q_j}}\right). \end{cases} \tag{4}$$

where X_{P,Q_j} (=0 or 1) denotes the occurrence of Q_j in a page P, TF_{P,Q_j} denotes the term frequency, the number of Q_js appear in a page P, and $TF_{P,max}$ denotes the maximum number of TF_{P,Q_j} through all terms Q_j $(j = 1, \ldots, M)$.

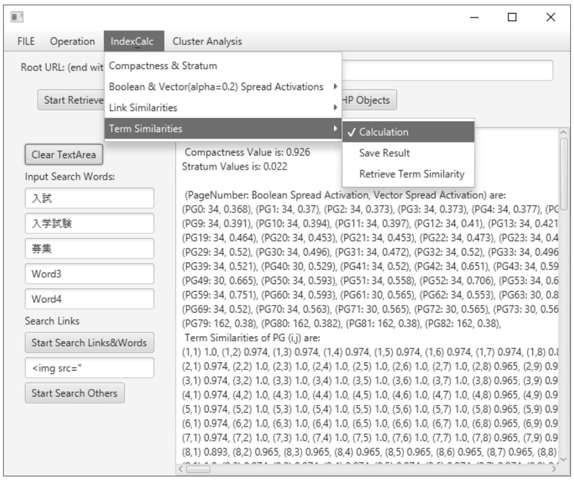
In order to measure a similarity of two pages P_i and P_j for a set of fixed query terms $q = \{Q_1, \cdots, Q_M\}$, the following index is defined as the term-based similarity index,

$$S^{terms}_{P_i,P_j} = \frac{\sum_{l=1}^{M} w_{P_i,Q_l} w_{P_j,Q_l}}{\sqrt{\sum_{l=1}^{M} w^2_{P_i,Q_l} \sum_{l=1}^{M} w^2_{P_j,Q_l}}} \tag{5}$$

where $w_{P_i,Q_l} = w^{TF}_{P_i,Q_l} w^{at}_{P_i,Q_l}$ with

$$w^{at}_{P,Q_j} = \begin{cases} 10 & \text{if } Q_j \text{ is in title of } P, \\ 5 & \text{if } Q_j \text{ is in headers or keywords} \\ & \text{or address in } P, \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

Since we did not implement a method to extract tags other than link tag from text, we set all the $w^{at}_{P,Q_j} = 1$. Figure 3 shows the calculation results of the term similarities.



```
 Compactness Value is: 0.926
Stratum Values is: 0.022

 (PageNumber: Boolean Spread Activation, Vector Spread Activation) are:
(PG0: 34, 0.368), (PG1: 34, 0.37), (PG2: 34, 0.373), (PG3: 34, 0.373), (PG4: 34, 0.377), (PG
(PG9: 34, 0.391), (PG10: 34, 0.394), (PG11: 34, 0.397), (PG12: 34, 0.41), (PG13: 34, 0.421
(PG19: 34, 0.464), (PG20: 34, 0.453), (PG21: 34, 0.453), (PG22: 34, 0.473), (PG23: 34, 0.4
(PG29: 34, 0.52), (PG30: 34, 0.496), (PG31: 34, 0.472), (PG32: 34, 0.52), (PG33: 34, 0.496
(PG39: 34, 0.521), (PG40: 30, 0.529), (PG41: 34, 0.52), (PG42: 34, 0.651), (PG43: 34, 0.59
(PG49: 30, 0.665), (PG50: 34, 0.593), (PG51: 34, 0.558), (PG52: 34, 0.706), (PG53: 34, 0.6
(PG59: 34, 0.751), (PG60: 34, 0.593), (PG61: 30, 0.565), (PG62: 34, 0.553), (PG63: 30, 0.8
(PG69: 34, 0.52), (PG70: 34, 0.563), (PG71: 30, 0.565), (PG72: 30, 0.565), (PG73: 30, 0.56
(PG79: 162, 0.38), (PG80: 162, 0.382), (PG81: 162, 0.38), (PG82: 162, 0.38),
 Term Similarities of PG (i,j) are:
(1,1) 1.0, (1,2) 0.974, (1,3) 0.974, (1,4) 0.974, (1,5) 0.974, (1,6) 0.974, (1,7) 0.974, (1,8) 0.
(2,1) 0.974, (2,2) 1.0, (2,3) 1.0, (2,4) 1.0, (2,5) 1.0, (2,6) 1.0, (2,7) 1.0, (2,8) 0.965, (2,9) 0.9
(3,1) 0.974, (3,2) 1.0, (3,3) 1.0, (3,4) 1.0, (3,5) 1.0, (3,6) 1.0, (3,7) 1.0, (3,8) 0.965, (3,9) 0.9
(4,1) 0.974, (4,2) 1.0, (4,3) 1.0, (4,4) 1.0, (4,5) 1.0, (4,6) 1.0, (4,7) 1.0, (4,8) 0.965, (4,9) 0.9
(5,1) 0.974, (5,2) 1.0, (5,3) 1.0, (5,4) 1.0, (5,5) 1.0, (5,6) 1.0, (5,7) 1.0, (5,8) 0.965, (5,9) 0.9
(6,1) 0.974, (6,2) 1.0, (6,3) 1.0, (6,4) 1.0, (6,5) 1.0, (6,6) 1.0, (6,7) 1.0, (6,8) 0.965, (6,9) 0.9
(7,1) 0.974, (7,2) 1.0, (7,3) 1.0, (7,4) 1.0, (7,5) 1.0, (7,6) 1.0, (7,7) 1.0, (7,8) 0.965, (7,9) 0.9
(8,1) 0.893, (8,2) 0.965, (8,3) 0.965, (8,4) 0.965, (8,5) 0.965, (8,6) 0.965, (8,7) 0.965, (8,8)
```

Figure 3: Term Similarity Calculation.

2.1.4 Clustering Method

We implemented two clustering algorithms such as the kernel k-means algorithm and "Structural Clustering Algorithm for Networks ($SCAN$)".

For a fixed number of clusters k and the kernel also denoted by k with feature map ϕ from a set of nodes G to the Hilbert space $\mathcal{H}$, the kernel k-means algorithm try to find a set of clusters $\{C_1, \ldots, C_k\}$ minimizing the following value.

$$\sum_{i=1}^{k} \sum_{x \in C_i} \| \phi(x) - \mu_i \|^2, \tag{7}$$

where

$$\begin{aligned}
\| \phi(x) - \mu_i \|^2 &= \| \phi(x) - \frac{1}{|C_i|} \sum_{x' \in C_i} \phi(x') \|^2 \\
&= k(x,x) - \frac{2}{|C_i|} \sum_{x' \in C_i} k(x,x') \\
&\quad + \frac{1}{|C_i|^2} \sum_{x'' \in C_i} \sum_{x' \in C_i} k(x'',x').
\end{aligned}$$

$SCAN$, proposed by Xiaowei Xu et al. (Xu et al., 2007), outputs three types of clusters such as "hub", "outlier", and ordinal clusters, by using structural similarity and two parameters $0 \le \varepsilon \le 1$ and $\mu \in \mathbb{Z}^+$.

2.1.5 Spectrum Analysis

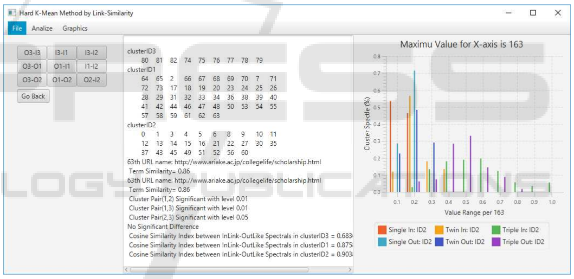
Shoda et al. considered all the connected sub-graph and calculate the total weight of each of them, then proposed to visually evaluate the similarity by graphing their frequency of appearance as the spectrum (Shoda et al., 2003).

For a weighted non-directed graph G and the set of weight $\{w(P)\}$, they consider all the connected subgraphs $\{SG \in 2^G; SG \text{ is connected}\}$ 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 the vector with component values of numbers of SGs whose weight are 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.

For a directed graph, we proposed the in-weight of connected subgraph SG as follows. Given 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 $w_{P_1 \to P_2} = w_1 w(P_1) + w_2 w(P_2)$ 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, $w_{direct}(\{P_1, P_2\})$ to be the average of $w_{P_1 \to P_2}$ and $w_{P_2 \to P_1}$. The out-weight is also defined in a symmetrical way.

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.

In the right pane of Figure 4, the spectrum charts corresponding to each cluster are shown.



```
clusterID3
     80   81   82   74   75   76   77   78   79
clusterID1
     64   65   2    66   67   68   69   70   7    71
     72   73   17   18   19   20   23   24   25   26
     28   29   31   32   33   34   36   38   39   40
     41   42   44   46   47   48   50   53   54   55
     57   58   59   61   62   63
clusterID2
     0    1    3    4    5    6    8    9    10   11
     12   13   14   15   16   21   22   27   30   35
     37   43   45   49   51   52   56   60
63th URL name: http://www.ariake.ac.jp/collegelife/scholarship.html
 Term Similarity= 0.86
63th URL name: http://www.ariake.ac.jp/collegelife/scholarship.html
 Term Similarity= 0.86
 Cluster Pair(1,2) Significant with level 0.01
 Cluster Pair(1,3) Significant with level 0.01
 Cluster Pair(2,3) Significant with level 0.05
No Significant Difference
 Cosine Similarity Index between InLink-OutLike Spectrals in clusterID3 = 0.683
 Cosine Similarity Index between InLink-OutLike Spectrals in clusterID1 = 0.875
 Cosine Similarity Index between InLink-OutLike Spectrals in clusterID2 = 0.903
```

Figure 4: Spectrum Chart in the right Pane.

3 IMPROVEMENTS

While the existing system has shown some performance, it leaves some room for improvement. In this section, we propose improvements in webpage retrieving within the website, link spectral evaluation for each in- and out-weight, and page similarity by keyword. We also describes a sentence related evaluation method that uses text readability on pages.

3.1 Webpage Retrieving

Instead of "URL" and "URLConnection" classes, we use "Jsoup" and "Elements" classes for retrieving and process the obtained page as a "Document" class object. The program below describes a epitome of the whole retrieving program, where "TAGS" is a set of

HTLM tags such as “<h1>”, “<h2>”, “<p>” etc.:

```
Document document=Jsoup.connect(url).get();
System.out.println(document.text());
for(String tag:TAGS) {
 ArrayList<String> strs=
   (ArrayList<String>)document.
     getElementsByTag(tag).eachText();
  System.out.println("<"+tag+">");
  for(String str:strs) {
     System.out.println(str);
  }
}
Elements links = document.select("a[href]");
```

By using these classes, we can improve the retrieval execution time and correct unexpected page access.

3.2 Link Spectral Evaluation

The graph structure obtained from the links between webpages is a directed graph, and it was necessary to define input link weights and output link weights for each node. Therefore, we defined the in-weight of a subgraph of only two nodes, $\{P_1, P_2\}$ as the weighted value $w_{P_1 \to P_2} = w_1 w(P_1) + w_2 w(P_2)$ with a fixed weights (w_1, w_2) such as $w_1 < w_2$ 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, $w_{direct}(\{P_1, P_2\})$ to be the average of $w_{P_1 \to P_2}$ and $w_{P_2 \to P_1}$.

In this definition, the weight value $w(P)$ is the number of link to or from any node of SG. Here we try to improve the index by changing this to hub weight or to authority weight according to out-link or in-link weight calculation. These weights, proposed by Kleinberg (Kleinberg, 1999), are importance values of each node as a referring or referred by other node in a directed graph. They are given as each component of the principal eigenvector (eigenvector for the greatest positive real eigenvalue) of the matrices AA^t and A^tA with the adjacent matrix A.

In a website, the number of nodes can be several hundred to several thousand, or sometimes exceeds 10,000, it is not easy or even impossible to implement codes obtaining principal eigenvetor in our system. Thus, we tried to use “Process” class to call and use an external computer algebra program such as *SageMath*[1] of *PARI/GP*[2]. Even if a computer algebra system is used, when the number of nodes is large and problems will occur in the output of the eigenvectors. Thus we will calculate for each cluster with about 50 nodes at most, then apply the system.

3.3 Term-based Index

In order to calculate and evaluate term-based indexes, we need to consider the differences in languages. Although almost sentences in webpages are written in English, there are many websites where some important information are represented only in language of each country. Thus, we add a language selection box and also add the menu item “Text Analysis” shown in the Figure 5.

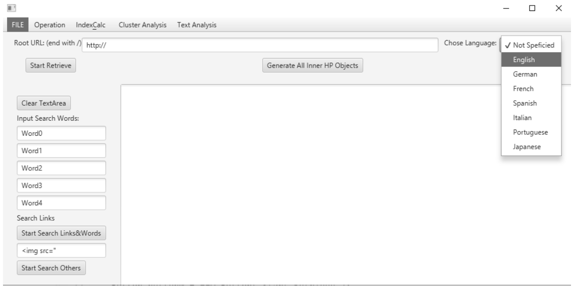


Figure 5: Language Selection Box and Text Analysis menu.

When we choose to the menu, new window for text analysis in the chosen language appears.

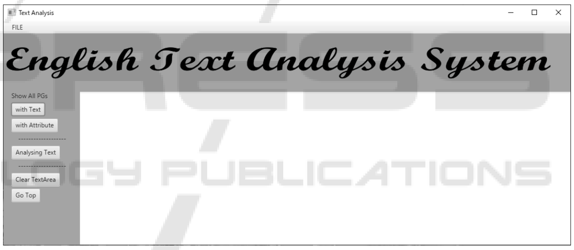


Figure 6: English Text Analysis Window.

In both the formulae of vector spread activation $RV_{P,q}$ and the term-based similarity index $S^{terms}_{P_i,P_j}$, it appears the term frequency TF_{P,Q_j} defined as in equation of (4). This value originally represented the number of times a word Q_j appeared on a webpage P, but some extended version have been used in a vector space model (Wong et al., 1985; Tsatsaronis and Panagiotopoulou, 2009; Waitelonis et al., 2015).

For a query word Q and a webpage P, we propose to extend TF_{P,Q_j} to be the summation of the number of appearance times the cosine similarity values $cos(W, Q_j)$ of word W throughout P. Considering the calculation cost, words are limited to specific particles such as noun, and only those with a certain degree of similarity or more are added. We also modify the attribution weights in (5) according to tag type in the

[1] https://www.sagemath.org/

[2] https://pari.math.u-bordeaux.fr/

following way.

$$w^{at}_{P,Q_j} = \begin{cases} 10 & \text{if } Q_j \text{ in <title>}, \\ 7 & \text{if } Q_j \text{ in <h1>} \\ 5 & \text{in <h2>}, \\ 3 & \text{in <b>, <i>, <u>}, \\ 1 & \text{otherwise}, \end{cases}$$

Then we extend the equation (6) by modifying the definitions of TF_{P,Q_j} and $\sum_{P'} X_{P',Q_j}$ in (4) as follows.

$$\begin{cases} TF_{P,Q} = \sum_{Q' \in Sim_{r_0}(P,Q)} cos(v_Q, v_{Q'}) \\ X_{P',Q} = max\{cos(v_Q, v_{Q'}); Q' \in Sim_{r_0}(P,Q)\}, \end{cases} \tag{8}$$

where v_Q denotes the word vectors corresponding to Q and $Sim_{r_0}(P,Q)$ is the set of all nouns Q' appear in the page P satisfying $cos(v_Q, v_{Q'}) \geq r_0$.

In order to implement the system above, some steps of natural language processing are required. At first, we need to extract words from text while decomposing them into particles, then examine their relationships. For this tasks, we use Tree-Tagger as morphological analysis application. The output of Tree-Tagger is the triplet of the word itself, the particle name, the prototype, and the symbols to represents them are different for each language.

For the calculation of the word vector v_Q, we apply FastText created by Tomas Mikolov which outputs the word distribution expression vector (the word vector) obtained by machine learning from a large corpus. There are files pre-learned from the Wikipedia dump files in many languages (157 languages) whose format is available to be executed by the fastText program.

3.4 Readability Index

As with term-based indexes, we will introduce readable indexes that correspond to each language.

To evaluate sentences in each webpage by their readability, we incorporate some of the following readability indexes with S, W, C, and Sy denoting the number of clauses, the number of words, the number of characters, and the number of syllables of the whole sentence, respectively.

- Flesch Reading Ease Score: A measure of text readability, proposed by Rudolf Flesh in 1946. (Flesch, 1946), (DuBay, 2004, p.21)

  $$206.830 - 1.015\frac{W}{S} - 84.6\frac{Sy}{W}$$

  – Table1 shows the correspondence between the score value, the difficulty level, and the genre of a typical book. Since it is an index showing readability, the lower the evaluation score, the harder it is to read.

  – In his 1978 dissertation "Wie verständlich sind unsere Zeitungen?", Toni Amstad in the Zürich University finds the formula for this index in German as follows. $180 - \frac{W}{S} - 58.5\frac{Sy}{W}$

  – It seems that the following formula (Huerta, 1959) by Fernández Huerta is often used as the corresponding evaluation index calculation formula in Spanish. $206.84 - 1.02\frac{W}{S} - 60\frac{Sy}{W}$

  – The table 2 shows the correspondence between the index values and the educational system in USA and Spain.

Table 1: Flesch Reading Ease Score and Readability Comparison.

Score	Difficulty	Representative Books
∼ 30	Very Difficult	Special Dissertation
30 ∼ 50	Difficult	General Academic Journals
50 ∼ 60	Somewhat Difficult	high quality magazine
60 ∼ 70	Standard	Article Summary
70 ∼ 80	Slightly Easy	Science Fiction
80 ∼ 90	Easy	Popular Novel
90 ∼ 100	Very Easy	Comic

Table 2: Country-specific Flesch Reading Ease Score and Grade Comparison.

Score	USA	Spain
∼ 30	University Graduate	University specialty
30 ∼ 50	Grades 13-16	Elective Course
50 ∼ 60	Grades 10-11	Before Entering University
60 ∼ 70	8-9 grades	7 and 8 grades
70 ∼ 80	Grade 7	Grade 6
80 ∼ 90	Grade 6	Grade 5
90 ∼ 100	5th Grade	4th Grade

- Dale-Chall Readability Score: In 1948, Edgar Dale and Jeanne Sternlicht Chall revised the Flesh Reading Ease Score and defined a readability index using an easy-to-read English word list of about 3,000 words called a "Dale-Chall's long list". (Dale and Chall, 1948), (DuBay, 2004, p.23)

  $$0.1579\frac{Dw}{W} + 0.0496\frac{W}{S} + 3.6365$$

  – Dw is the number of words not found in "Dale-Chall's long list".

  – The table 3 gives the correspondence between the evaluation values and grades in USA.

- Fog Index: Robert Gunning states that "many of the reading problems are writing problems, and newspapers and corporate documents are covered in Fog and filled with unnecessary complexity". (DuBay, 2004, p.25). In the book "Writing Technique of Clear Writing" (Gunning, 1952), he pro-

Table 3: Dale-Chall Grade Evaluation.

Score	Corresponding Grade Level (in USA)
~ 4.9	4th Grade or Younger
$5.0 \sim 5.9$	5th and 6th Grade
$6.0 \sim 6.9$	7th and 8th Grade
$7.0 \sim 7.9$	9th and 10th Grade
$8.0 \sim 8.9$	11th and 12th Grade
$9.0 \sim 9.9$	13th, 14th and 15th Grade
$10.0 \sim$	16th Grade or Above (University graduation)

posed the readability index for adults called the "Fog Index".

$$0.4(\frac{W}{S}+100\frac{Sy^+}{W}),$$

where Sy^+ represents the number of words with 3 or more syllables.

- Flesh-Kincaid Readability Grade Level: An index to measure the difficulty of reading a sentence.(DuBay, 2004, p. 50)

  $$0.39\frac{W}{S}+11.8\frac{Sy}{W}-15.59$$

  - There is no linear dependency between this index and the aforementioned Flesh Reading Ease Score.
  - Table 4 represents an index to find the value corresponding to the education level in the United States.

Table 4: Flesh-Kincaid Readability Grade Level Evaluation.

Reading level	rating	book example
Basic	$0 \sim 3$	Preparation for Reading
	$3 \sim 6$	Picture Book
Average	$6 \sim 9$	Harry Potter
	$9 \sim 12$	Jurassic Park
Mastery	$12 \sim 15$	Historical Novel
	$15 \sim 18$	Academic Papers

- SMOG (Simple Measure Of Gobbledygook) Grading: While many indicators are given in linear form of the number of characters, the number of syllables, etc., Gobbledygook Harry McLaughin proposed non-linear formula for evaluating the level of education for understanding sentences in the UK. (McLaughlin, 1969), (DuBay, 2004, p.47)

  $$3.1291+1.0430\sqrt{30\frac{Sy^+}{S}}$$

  - Table5 correspondence table with education level in the UK.

Table 5: Correspondence between SMOG Grading Index Value and Education Level.

Evaluation Value	Education Level
~ 4.9	Elementary School
$5.0 \sim 8.9$	Secondary School
$9.0 \sim 12.9$	High School
$13.0 \sim 16.9$	Undergraduate
17.0 and above	College Graduate

  - G-SMOG by Bamberger-Vanecek is the following equation for German.(Bamberger and Vanecek, 1984) $\sqrt{30\frac{Sy^+}{S}-2}$
  - SOL is the following formula for Spanish and French.(Contreras et al., 1999)

    $$(3.1291+1.0430\sqrt{30\frac{Sy^+}{S}})\times 0.74-2.51$$

4 CONCLUSIONS

We have proposed the following improvements to the existing website evaluation system we have developed.

- By using Java Jsoup class, the stability of all the webpage retrieving in the website has increased.
- By making good use of computer algebra programs, it is possible to perform more precise spectral analysis of directed graphs created by link structures.
- By incorporating morphological analysis and word dispersion expression vector as a system of natural language processing, more flexible keyword search can be performed.
- By implementing sentence readability index, It is possible to evaluate the readability bias of text expressions on webpages.
- Made the entire process compatible with several languages.

Our future task is to apply the improved program and evaluate some websites to verify its effectiveness.

REFERENCES

Bamberger, R. and Vanecek, E. (1984). *Lesen-Verstehen-Lernen-Schreiben: Die Schwierigkeitsstufen von Texten in deutscher Sprache*. Diesterweg-Sauerländer.

Chen, C. (1998). Generalised similarity analysis and pathfinder network scaling. *Interacting with Computers*, 10(2):107–128. DOI: 10.1016/s0953-5438(98)00015-0.

Cherl, N. M. and Locsin, R. J. F. (2018). Neural networks application for water distribution demand-driven decision support system. *Journal of Advances in Technology and Engineering Studies*, 4(4):162–175. DOI: 10.20474/jater-4.4.3.

Contreras, A., G-A., R., E., M., and D-C., F. (1999). The sol formulas for converting smog readability scores between health education materials written in spanish, english, and french. *Journal of Health Communications*, 4:21–29.

Dale, E. and Chall, S. (1948). A formula for predicting readability. *Educational Research Bulletin*, 27(1):1–20.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271. DOI: 10.1007/bf01386390.

DuBay, W. H. (2004). The principles of readability. In *Impact Information*, pages –, Cost Mesa California.

Flesch, R. (1946). *The art of plain talk*. Harper, New York.

Floyd, R. W. (1962). Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345–350. DOI: 10.1145/367766.368168.

Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

Huerta, F. J. (1959). Medidas sencillas de lecturabilidad. In *Revista pedagógica de la sección femenina de Falange ET y de las JONS*, pages 29–32.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632. DOI: 10.1145/324133.324140.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). Aimq: A methodology for information quality assessment. *Information & Management*, 40(2):133–146. DOI: 10.1016/s0378-7206(02)00043-5.

Liang, G. and Nagata, K. (2011). A study on e-business website evaluation formula with variables of information quality score. In *Proceedings of the 12th Asia Pacific Industrial Engineering and Management Systems Conference*, pages –.

McLaughlin, G. H. (1969). Smog grading-a new readability formula. *Journal of Reading*, 12(8):639–646.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *ICLR Workshop Paper*, pages –.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2015). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Nagata, K. (2019). Website evaluation using cluster structures. *Journal of Advances in Technology and Engineering Research*, 5(1):25–36. DOI: 10.20474/jater-5.1.3.

Noorzad, A. N. and Sato, T. (2017). Multi-criteria fuzzy-based handover decision system for heterogeneous wireless networks. *International Journal of Technology and Engineering Studies*, 3(4):159–168. DOI: 10.20469/ijtes.3.40004-4.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620. DOI: 10.1109/icectech.2011.5941988.

Schvaneneldt, R. W., Dearholt, D., and Durso, F. (1988). Graph theoretic foundations of pathfinder networks. *Computers & Mathematics with Applications*, 15(4):337–345. DOI: 10.1016/0898-1221(88)90221-0.

Shoda, R., Matsuda, T., Yoshida, T., Motoda, H., and Washio, T. (2003). Graph clustering with structure similarity. In *Proceedings of the 17th Annual Conference of the Japanese Society for Artificial Intelligence*, pages –, New York, NY.

Thorup, M. (2004). Integer priority queues with decrease key in constant time and the single source shortest paths problem. *Journal of Computer and System Sciences*, 69(3):330–353. DOI: 10.1016/j.jcss.2004.04.003.

Tsatsaronis, G. and Panagiotopoulou, V. (2009). A generalized vector space model for text retrieval based on semantic relatedness. In *Proceeding of EACL 2009 Student Research Workshop*, pages 70–78, Athens, Greece. Association for Computational Linguistics.

Waitelonis, J., Exeler, C., and Sack, H. (2015). Linked data enabled generalized vector space model to improve document retrieval. In *CEUR Workshop Proceedings*, pages –. CEUR-WS.org.

Warshall, S. (1962). A theorem on boolean matrices. In *Proceedings of the ACM*, pages –, Berlin, Germany. ISSN 1613-0073.

Weiss, R., Velez, B., Sheldon, M., Namprempre, C., Szilagyi, P., Duda, A., and Gifford, A. (1996). Hypursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of the 7th ACM Conference on Hypertext*, pages 180–193.

Wong, S. K. M., Ziarko, W., and Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceeding of the 8th SIGIR Conference on Research and Development in Information Retrieval*, pages 18–25. ACM.

Xu, X., Yuruk, N., Feng, Z., and Schweiger, T. A. (2007). Scan: A structural clustering algorithm for networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages –, Jakarta, Indonesia. ISSN:2414-.