Social Network Analysis on a Topic-Based Navigation Guidance System in a Public Health Portal

Jin Zhang
School of Information Studies, University of Wisconsin Milwaukee, 2025 E Newport Ave., Milwaukee, WI 53211. E-mail: jzhang@uwm.edu

Shanshan Zhai
School of Information Management, Central China Normal University, No.152, Luoyu Street, Wuhan City, Hu Bei Province 430079, China. E-mail: kaede_ss@163.com

Hongxia Liu
School of Information Resource Management, Ren Min University of China, No. 59, Zhongguancun Street, Haidian District, Beijing 100872, China. Email: elise2012@126.com

Jennifer Ann Stevenson
School of Information Studies, University of Wisconsin Milwaukee, Milwaukee, WI 53211. E-mail: steven58@uwm.edu

We investigated a topic-based navigation guidance system in the World Health Organization portal, compared the link connection network and the semantic connection network derived from the guidance system, analyzed the characteristics of the 2 networks from the perspective of the node centrality (in_closeness, out_closeness, betweenness, in_degree, and out_degree), and provided the suggestions to optimize and enhance the topic-based navigation guidance system. A mixed research method that combines the social network analysis method, clustering analysis method, and inferential analysis methods was used. The clustering analysis results of the link connection network were quite different from those of the semantic connection network. There were significant differences between the link connection network and the semantic network in terms of density and centrality. Inferential analysis results show that there were no strong correlations between the centrality of a node and its topic information characteristics. Suggestions for enhancing the navigation guidance system are discussed in detail. Future research directions, such as application of the same research method presented in this study to other similar public health portals, are also included.

Introduction

Public health portals provide a variety of common health users with important health information channels. Public health portals, such as the WHO (World Health Organization, 2014) portal, are usually maintained by government agencies or professional health organizations. They offer rich, diverse, and authoritative health information for users. For instance, the WHO portal posts updates, risk assessments, technical guidance, travel and health advice, and frequently asked questions on the recent Middle East Respiratory Syndrome Coronavirus outbreak so that health users can access the much-needed information in a timely fashion. Many public health portals have built-in retrieval mechanisms with which people can search information, browse web pages of interest, and navigate the information space. These search means include, but are not limited to, internal search engines, sitemaps, subject directories, topic indices, and so on. In this study, the WHO portal was selected and investigated because it is an authoritative, reliable, and popular public health portal, has rich information, and maintains a subject-oriented navigation guidance system. The topics and associated related link relationships in the guidance system can generate a topic network for investigation.

Social network analysis examines, measures, and evaluates the characteristics of a network and connections among its actors (or nodes) in the network. Actors can be people,
organizations, or components of an investigated entity. Social network analysis focuses on the interactions between the actors. The relationships, which result from the interactions, can be information exchanged between the actors, a variety of physical or abstract connections between the actors such as hyperlinks in a web page or friendships in a community. Social network analysis has proven its merits in a wide range of application domains (Burt, 2000; Cook & Whitmeyer, 1992; Kempe, Kleinberg, & Tardos, 2003; Park, 2003). It provides a powerful, intuitive, and vivid visual analysis means for network data.

The primary aims of this study are to investigate a topic-based navigation guidance system in a public health portal (the WHO portal, 2014), compare the link connection network and the semantic connection network generated from the navigation guidance system, analyze the correlations between the centralities of a node and its information characteristics in the navigation guidance system, and, finally, provide constructive suggestions to optimize and enhance the topic-based navigation guidance system.

The originality of the study is threefold: ascertain a topic-based navigation guidance system, which is widely used in a web portal; apply a mixed research design of the social network analysis method, clustering analysis method, and inferential statistical analysis methods; and evaluate the navigation guidance system from both the related link relationships and the semantic content relationships’ perspectives.

The study’s findings can be used to optimize and enhance the structures of the navigation guidance system in three ways: (a) adding new related topics to a specific topic; (b) recommending potential topics that need to enrich their general information, technical information, publication, and multimedia categories; and (c) establishing a sophisticated topic-oriented hierarchical navigation guidance system. In addition, the study provides a new unique research method for the evaluation of similar navigation guidance systems in other web portals.

**Related Work**

**Public Health Online Information Needs**

People turn to the Internet to find reliable sources of information. The increasing number of online information health portals attests to this trend (Benigeri & Pluye, 2003; Evers, 2006). It is important that people are able to find reliable health information online (Cline and Haynes, 2001).

The widespread adoption adaptation of online health searching comes a need to evaluate criteria for health website evaluation such as accessibility, content, credibility, and scope (Benigeri & Pluye, 2003). Benigeri and Pluye (2003) focused on the shortcomings of health information websites and noted that in order to minimize the possibility of bad health websites it was of the utmost importance that health professionals be involved with the design and evaluation of online information systems. In relation to health website evaluation, Zhang, Wolfram, Wang, Hong, and Gillis (2008) found significant differences among health query terms and medical terminology.

According to a survey conducted by Pew Internet, 35% of Americans use online research to figure out a medical condition (Fox & Duggan, 2013). Perhaps even more interesting is that 46% of those who used the Internet to conduct a medical search concluded that they should see a medical professional. Even outside of a health information portal, a great deal of information is sought within an online community.

**Web Portal Navigation and Search Aids**

For a web portal, information search mechanisms are critical to its success and ease of access for users. There are two main methods for users to interact with a web health portal during information retrieval: query searching and browsing. The design of the interface is also critical because of the connection it gives between the users and the information.

Many case studies have been conducted in relation to health information portals and design usability (Borgatti & Foster, 2003; Griffiths et al., 2012; Kwiat, Valente, & Celentano, 2001; Yousefi-Nooraie, Dobbins, Brouwers, & Wakefield, 2012). It seems evident that health information portals and their online search capabilities will have to change in order to comply with user needs and expectations.

Health information is often organized into subject directories, which are an efficient means for organizing information. Many studies conducted on subject directories automation and maintenance in the web environment have evaluated different organization methods. Yang and Lee (2004) applied a text-mining technique based on SOM (self-organizing map) to automatically create web directories and organize web pages of the data used into hierarchies. Users’ transaction data were used to optimize the health subject directory by using the SOM method and component plane analysis method (Zhang & An, 2010; Zhang, An, Tang, & Hong, 2009). The component plane method revealed and demonstrated the distribution of a specified component (attribute) of interest in a categorized fashion and the method used as a tool in exploratory data analysis. Subsequently, Yang and Lee (2006) came up with NaviSOM in support of a navigational structure for World Wide Web information.

**Social Network Analysis**

Social network analysis (SNA) reveals structural relations between entities or nodes. SNA has its origins in a study (Coleman, 1986) on theories of social action. Barnes (1954) first theorized the concept of SNA. The strength of SNA comes from its potential for an immense system that can be applied to many domains.

Tichy, Tushman, and Fombrun (1979) recommended that organizations utilize the theoretical structure to help identify patterns and conceptualizations within the organization.
Scott (2000) described the relations between nodes as the social positions or relations within the social network. Clustering analysis revealed that the level of connectedness varied among the following variables: Sharing options, seeking advice, influence of advice, receiving advice, and explicit support of brands was obvious.

The relationships and connections that emerge through the application of SNA can be revealing. SNA studies relationships in terms of nodes and ties, with nodes being the individual entities in the social network and ties being the relationships between individual entities. Research can be conducted in a number of different ways, depending upon whether the entities and relationships are abstract or physical objects. For instance, a physical object can be a web page and an abstract object can be a cluster of keywords or documents in a database. Salah, Manovich, Salah, and Chow (2013) analyzed user-generated content that had been applied to a number of different media (i.e., videos, games, and images). These were the nodes. The relationships revealed through SNA can shed light into areas that are very subject specific. Hambrick (2012) used SNA to explore sport social networks, and found that through the integration of social network analysis, those involved in the sports industry promote products through Twitter. Hambrick (2012) was able to illuminate the interworking of athletes, teams, and leagues through the use of Twitter.

Guo (2012) applied SNA as a framework for media use and public agenda to provide a better understanding of news coverage. Guo was able to identify the important nodes within the network of news coverage. Rolland and Parmentier (2013), who took the concept of bulletin boards as a concept of communication and applied to modern-day social media, and then compared both concepts within a SNA framework, explored another avenue. Rolland and Parmentier (2013) discovered that the power found in social media applications was key owing to the continuum of communication and additional networks nodes and that were established over time.

The integration of SNA and community associations can reveal not only the linkages of the network, but also address semantic associations (Zhao et al., 2012). Zhao et al. (2012) used the subspace clustering algorithms to draw out the topics shared by people in the network, which detects all potential clusters in a network. Once the nodes within the social network have been identified, the next step is to analyze the relationships between those nodes. For example, the identification of the nodes as the athletes, teams, and leagues permitted Hambrick (2012) to dive deeper into their relationships in Twitter. This enabled the researchers to discover how information was spreading through the network, for example, the use of promotional tweets made by race organizers to provide information to followers. Hoppe and Reinelt (2010) discussed the importance of identifying strong and weak ties within leadership development and were able to break down different social networks as a result of identifying the relationships between the nodes—for example, peer leadership network, organizational leadership network, field-policy leadership network, and collective leadership network (Hoppe & Reinelt, 2010, p. 601). Measurement of the strong and weak ties can be done in a number of ways, depending upon the social network. Grabowicz, Ramasco, Moro, Pujol, and Eguiluz (2012) found that the more mentions exchanged between users, the stronger the ties between them. Knoke and Yang (2008) use the term cliques to describe the cohesion or relations between those involved in a social network. There are also strong identifiers associated with each clique. The connectivity among groups within social media is very complex.

Chambers, Wilson, Thompson, and Harden (2012) wanted to improve the effectiveness of the decision-making process within health care settings. They used social network analysis to discover the types of relationships between health care coworkers and were able to identify certain thematic ties, for instance, social influence, describing behavior, and diffusion of motivations (Chambers et al., 2012, p. 7).

SNA and public health information seeking are two areas that, when combined, can cast light into health care settings, and there have been several studies that have already noted the importance of the combination. Interestingly, SNA has been applied to a number of other areas within public health, such as teenage cigarette smoking habits and family communication and heredity diseases (Koehly et al., 2003; Valente, Unger, & Johnson, 2005). Luke and Harris (2007) discussed the history of social network analysis and how it was utilized to better understand disease transmission in HIV/AIDS and other sexually transmitted diseases. They concluded that the organizational structure produced through SNA would be a very effective tool when combined with public health communication methods. In addition, Scott et al. (2005) used SNA to compare communication patterns in primary care practices. Scott et al. (2005) found, in their analysis, that SNA could be used to help health organizations with organizational change and other decision-making communication exchanges.

People often communicate with one another to understand health concerns. In online setting, health concerns are discussed through health information portals. As a result, it is extremely important to ensure that the portals provide accurate, reliable, complete, and well-organized information so that users can quickly find information they need. For example, Mertens, Saint-Charles, and Mergler (2012) found through social communication network analysis that the men and women of a fishing village in the Amazon communicated through personal networks to provide information regarding mercury issues occurring in the fish.

Smith and Christakis (2008) explored how the interconnectedness of our culture has an impact on people’s health, and argued that health search mechanisms need to draw on SNA in order to provide people with new ways to manage their health care. The interactivity of public health online portals makes them a viable source of information for the public; however, they have to provide users with relevant information that is useful, pertinent, and accurate.
Gray, Klein, Noyce, Sesselberg, and Cantrill (2005) explored adolescent perceptions of health information and the Internet and found that adolescents use the Internet to explore health concerns on a regular basis, but were typically unsure of the trustworthiness of the information they found online. Pow, Gayen, Elliott, and Raeside (2012) applied social network analysis to nursing research and were able to identify relevant individuals or actors who influenced social norms and communication, which could be critical for services providers to have a deeper understanding of the network. Bossche and Segers (2013) applied SNA to the transfer of training as a way to investigate how people communicate and connect within a large web structure. Bossche and Segers (2013) found that SNA was able to add workplace perspectives that were not previously revealed, such as the role of the social network as an organization, the network outside the organization. Zhang, Yu, Fan, and Duan (2013) discussed emerging issues in collaborative health management. They discovered that there was a gap in communication between different institutions, and noted that health management research should be enhanced through the collaboration and internationalization of health management.

**Summary**

There remains the a need to evaluate online resources, including health information portals. Chambers et al. (2012) found that there is little awareness of SNA in health care settings. Health information portals have different purposes there are news sources, such as WebMD, and health information sites, such as WHO. One of the best ways to evaluate these navigation guidance systems in a public health portal is through social network analysis. Fortunately, the framework of SNA provides the theory and means needed to conduct the research. Few research studies on evaluation of a navigation guidance system in a public health portal using an SNA method have been conducted.

**Research Methods**

**Introduction**

The WHO portal was selected for this research study because it is a reputable and reliable public health website and the portal incorporates a topic navigation guidance system, which assists people in browsing and finding the right information in the portal. In addition, the size of the topics (207) in the system is reasonable for data processing and analysis. Each topic is associated with a web page covering the information on that subject topic and four information categories (general information, technical information, publications, and multimedia). Each topic includes a group of related topics. A topic in bold in Figure 1 indicates that it has a related topic. Related topics are provided in the navigation guidance system for browsing related information to the topic. All subject topics are alphabetically organized and presented in its entry web page under health topics (Figure 1).

The relationship between a topic and one of its related topics is directional. If all the subject topics in the navigation guidance system are extracted and the relationships among these subject topics are established by the provided related connections among the topics in the portal, then the identified topics can form a directional topic network. As a result, the network provides an ideal environment in which SNA methods and techniques can be applied.

After selecting WHO as the central website for this study, all the topics were extracted from the navigation guidance system and each of the subject topics was considered as a node in the network. A topic node $S_i$ is linked to another topic node $S_j$ in the subject topic set if the subject topic $S_j$ is listed as one of the related subjects/topics/links to the subject topic $S_i$. If a subject is regarded as a related subject topic of another subject topic, it means that they are related and therefore they should be connected in the network. This kind of relationship between the two topics indicates their semantic and conceptual associations and connections. In other words, a topic node $S_i$ is linked to another topic node $S_j$ in the network if the subject topic $S_i$ is listed as one of the related subjects/topics in the topic web page $S_j$. As a result, a connection between the topic node $S_i$ and the topic node $S_j$ was established in the network. After all the topic nodes in the navigation guidance system were processed, a topic link network was formed based on inbound link relationships of the topic nodes.

However, at this point, the connection between the subject node $S_i$ and the subject node $S_j$ in the topic link network does not accurately indicate the semantic strength between the subject node $S_i$ and the subject node $S_j$. Notice that the degree to which the two topics are semantically or conceptually associated may vary in different contexts, depending on the topic content. If the network analysis is only based on the inbound link connections among the topics, then the results may not fully reflect the semantic relationships among the topics, which are critical and vital for the navigation guidance system. To solve this problem, we can factor in the similarity between the two linked topics. That is, the similarity between the two subject topics was computed and assigned to the connection between the two topic nodes as a weight in the network. In this way, both an inbound link connection of the two topics and their semantic strength of the connection were considered in the link connection network.

Each subject topic was associated with a web page. The association includes detailed information on the topic in the navigation guidance system. Regardless of the related connections (links) among the topics in the navigation guidance system, the semantic strength between any two topics in the topic set can be determined and calculated based on the corresponding web page contents of the two topics, rather than the related link connections. If the semantic strength between two topics is employed as an abstract connection between them, all the topic nodes involved in the navigation guidance system can generate another semantic connection network. This semantic connection network is quite different from the link connection network in terms of the node relationships, although the two networks share the same relationships.
topic nodes. These two networks play an extremely important role in this study. When the link connection network was evaluated, the semantic connection network served as a baseline for the evaluation analysis.

The application of SNA methods to both the semantic connection network and the link connection network, respectively, the network characteristics of the semantic connection network, and the link connection network were analyzed and compared. If the SNA results demonstrate the network characteristics of both the semantic connection network and the link connection network are similar, it implies that the link connections among the topics in the navigation guidance system correctly reflect the semantic connections of the topics. In other words, the navigation guidance system is sound. Otherwise, the navigation guidance system needs to be optimized and enhanced. If the related link connections among the topics in the navigation guidance system cannot correctly reflect the semantic connections of the topics, the system would not effectively guide users to right places during navigation.

**Data Preparation, Processing, and Organization**

Having identified a public health portal with a topic-based navigation guidance system where each subject topic maintains a group of related topic links, all subject topics and their related topics in the portal were traversed, all the topic web pages were kept, and keywords in each of the corresponding topic web pages were extracted to form a keyword set.

A keyword validation process followed. A stop-word list was used to filter useless words in the extracted keyword set. These stop-words include grammar function words and other words bearing no retrieval meanings, such as “of,” “in,” “with,” “a,” “an,” “the” . . . Synonyms were combined, for example, heart disease was replaced by cardiac disease. Multiple forms of a regular term were converted to their regular term (e.g., “diseases” was converted to “disease”).

Following the keyword validation process, a topic-keyword matrix (TKM) was created. The rows of the matrix are topics and the columns of the matrix are keywords that appear on the corresponding topic web pages. In other words, the keywords were directly extracted from the subject topic web pages to describe the topics in the matrix. A cell \(c_{ij}\) of the matrix indicates the degree to which a topic \(i\) and the corresponding keyword \(j\) are related. The degree is also called the weight, which is primarily determined by its inverted document frequency value in the collected data set. The matrix is described in Equation (1). Here, \(n\) is the

![Image](http://www.who.int/topics/en/)
number of the topics whereas $m$ is the number of the unique keywords extracted from the entire topic web pages in the navigation guidance system.

\[
TKM = \begin{bmatrix}
c_{i1} & \cdots & c_{im} \\
\vdots & \ddots & \vdots \\
c_{n1} & \cdots & c_{nm}
\end{bmatrix}
\]

(1)

A topic-topic inbound link matrix ($TTILM$) is defined. It is used to describe the inbound link relationship among the topics. A cell value of the matrix ($l_{ij}$) is not binary. Here, $n$ is the number of all the topics identified in the navigation guidance system. If the subject topic $i$ is linked by another subject topic $j$ (or the subject topic $j$ links to the subject topic $i$), then 1 is added to $l_{ij}$. Otherwise, 0 is added to $l_{ij}$. It is clear that the $TTILM$ is an $n \times n$ asymmetrical matrix because there are $n$ different subject topics, and that one subject topic $i$ linked to another subject topic $j$ does not ensure the subject topic $j$ linked to the subject topic $i$ (Equation [2]).

\[
TTILM = \begin{bmatrix}
l_{i1} & \cdots & l_{in} \\
\vdots & \ddots & \vdots \\
l_{n1} & \cdots & l_{nn}
\end{bmatrix}
\]

(2)

It is apparent that the following equation always holds because a subject topic can be considered as being linked by itself (Equation [3]).

\[l_{ii} = 1, \quad 1 \leq i \leq n\]

(3)

Based on the two matrices $TKM$ and $TTILM$, a new topic-topic weighed inbound link matrix ($TTWILM$) was generated. The $TTILM$ has the same column-row structure as the $TTWILM$. But the cell definitions of the two matrices are different (Equation [4]).

\[
TTWILM = \begin{bmatrix}
w_{i1} & \cdots & w_{in} \\
\vdots & \ddots & \vdots \\
w_{n1} & \cdots & w_{nm}
\end{bmatrix}
\]

Here, $w_{ij}$ is defined as the follows. Equation (5) shows that the weight $w_{ij}$ between two subject topics $i$ and $j$ is determined by their corresponding vectors defined in the $TKM$ (Equation [1]), because we know that a vector corresponds to a subject topic in the $TKM$. If a cell value ($l_{ij}$) in the $TTILM$ is equal to 0, then its counterpart $w_{ij}$ in the $TTWILM$ is also equal to 0. That is because if a cell value ($l_{ij}$) in the $TTILM$ is equal to 0, there is no connection between the two topic nodes $i$ and $j$. Therefore, there is no need to calculate their weight or connection strength.

We also have (Equation [6]):

In fact, the weight of the two topics $i$ and $j$ was calculated by a similarity measure. There are many similarity measures available and each has its strength and weakness. The similarity measure used in Equation (5) is called the cosine similarity measure. The cosine similarity measure (Korfhage, 1997) can effectively identify documents with the same keyword distribution and proportional keyword weights, and it has been widely used in information retrieval (Zhang & Rasmussen, 2001). According to Equations (5) and (6), the $TTWILM$ is also an $n \times n$ asymmetrical matrix like the $TTILM$.

Notice that in the portal regular web pages, in addition to the topic web pages can also be listed as related websites of a subject topic. Each of these related non-topic websites is associated with a specific subject topic. That is, two related subject topics are connected by a non-topic website in the portal. For this reason, these two subject topics maintain a connection in the $TTILM$.

If two subject topics have multiple connections in the navigation guidance system, the corresponding weight or connection strength in the $TTWILM$ increases accordingly.

The matrix $TTWILM$ indicates not only an inbound link connection between two topics, but also their connection strength (weight). In fact, the $TTWILM$ describes the link connection network discussed before. It is apparent that the network determined by the $TTWILM$ is directional.

Finally, the subject semantic matrix ($SSM$) used to describe the semantic relationships among the topics is defined as follows (Equation [7]):

\[
SSM = \begin{bmatrix}
t_{i1} & \cdots & t_{in} \\
\vdots & \ddots & \vdots \\
t_{n1} & \cdots & t_{nn}
\end{bmatrix}
\]

(7)

\[
t_{ij} = \frac{\sum_{k=1}^{n} c_{ik} \times c_{jk}}{\left(\sum_{k=1}^{n} c_{ik}^{2} \times \sum_{k=1}^{n} c_{jk}^{2}\right)^{1/2}}
\]

(8)

where the $SSM$ is a symmetric matrix because the similarity between two topics $T1$ and $T2$ is the same as that between two topics $T2$ and $T1$. In fact, Equation (8) attests to that. The subject semantic matrix describes the semantic connection network. It is clear that the network determined by the $SSM$ is not directional.
An adjusted subject semantic matrix (ASSM) is introduced because the original SSM was directly generated from the collected raw data, and there were many cells with low values. They made little contribution to the later analysis and resulted in overwhelming connections in the visual social network displays. Therefore, the cells with low values were set to 0 in the ASSM. Here, \( t'_{ij} \) is a cell in the ASSM. The ASSM is different from the SSM because some cells in the ASSM are set to 0 if the following condition described in Equation (11) is satisfied. A cut-off point is defined in Equation (9), if the value of a cell in the ASSM is smaller than the cut-off point; the cell is set to 0 (Equations [9]–[11]).

\[
CutOffPoint = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} t_{ij}}{n \times n} \tag{9}
\]

\[
ASSM = \begin{bmatrix}
    t'_{11} & \cdots & t'_{1n} \\
    \vdots & \ddots & \vdots \\
    t'_{n1} & \cdots & t'_{nn}
\end{bmatrix} \tag{10}
\]

\[
t'_{ij} = \begin{cases}
    t_{ij}, & t_{ij} \geq CutOffPoint \\
    0, & t_{ij} < CutOffPoint
\end{cases} \tag{11}
\]

It is clear that either the SSM or the ASSM is an \( n \times n \) symmetrical matrix unlike the TTILM and TTWILM. The ASSM describes the adjusted semantic connection network.

The SSM was used for the node clustering analysis whereas the ASSM was designed for network centrality analysis, which is discussed later. Both the networks determined by these two matrices are undirected.

The SSM and the ASSM play an important role in describing semantic relationships among the involved subject topics in the navigation guidance system. Both the matrices served as a baseline for the later comparison analysis between the semantic connection network and the link connection network in the portal.

Network Comparison Analysis

In order to investigate the validity and quality of the topic-based navigation guidance system in the portal, we needed to examine whether the link connection network was consistent with the semantic connection network by using the SNA method. It has a twofold process. First is clustering result comparison. If the resultant clusters from the link connection network are similar to those from the semantic connection network, it implies that the navigation guidance system is sound. Otherwise, the navigation system is not well designed and needs improvement. The second one is network node feature comparison. The network node feature refers to density and centrality of a social network. The density is defined as the number of all actual edges (or connections) divided by the number of all possible edges in a network. It indicates the connection strength of a network.

The centrality refers to the indicators that identify the most important nodes in a social network. These indicators were introduced later. If the network node feature data from the link connection network are not significantly different from those from the semantic connection network, it suggests that the validity of the navigation guidance system is confirmed. Otherwise, the navigation guidance system needs to be optimized. The detailed concepts and methods of these two ways are discussed separately.

The convergence of iterated correlation (CONCOR) method is one of the block models in SNA. It can be used to partition a social network into several social subnetworks based on node correlations in the social network (Schwartz, 1977). Each social subnetwork is a block. In fact, the CONCOR method is an effective clustering algorithm used to partition nodes in a network.

It was used to calculate the correlation between two nodes/topics based on structural equivalence in the matrix TTWILM. It results in a correlation matrix. Nodes with similar connection structures receive the highest correlation score of 1. Nodes with dissimilar connection structures receive the lowest correlation score of -1. The process repeats for resultant subnetworks until the cells of the correlation matrix converge. Then, the CONCOR algorithm permutes the nodes/topics in the correlation matrix into several distinct blocks. Finally, the nodes that share the similarity are clustered together in a block.

There are two salient characteristics of the CONCOR method. One is that the resultant blocks are generated based on the node connections in a social network. The second is that there is not any overlapping part among the resultant blocks or subnetworks. In other words, a node belongs to one subnetwork and to only one subnetwork. These characteristics are important because it makes it possible to compare the clustering results of the link connections of the link connection network with the clustering results from the semantic connections of the semantic connection network.

A pilot study was conducted to ascertain whether or not the clan analysis method was suitable for this study. The results show that the clan analysis method did not fit this study because of overlapping items among the resultant clans.

After the CONCOR method was applied to the TTWILM, \( k \)-blocks were produced.

Then, the \( k \)-means clustering method was employed to the SSM to yield \( k \)-clusters from the semantic connection network. It is apparent that the SSM serves as the input data set for this purpose given that the SSM describes the semantic connections of the topics. The parameter \( k \) is the number of the blocks/clusters in the CONCOR analysis. Owing to the fact that the two clustering analyses resulted in the same number of the clusters, it makes the clustering result comparison fair.

After the two clustering analyses from the two different methods, the \( F \)-value analysis, a popular measure for evaluating effectiveness of a cluster algorithm (Larsen &...
Aone, 1999; van Rijsbergen, 1979), was conducted to measure the similarity between the two clustering analysis result sets. The F-value measure can evaluate the similarity between two clustering sets from two different clustering analysis methods. Assume \( C1 \) represents a cluster from the topic link connection analysis \( S1 \) with a size of \( n_c \), \( C2 \) represents a cluster from the topic semantic connection analysis \( S2 \) with a size of \( n_s \) and \( n_t \) represents the number of the nodes/topics in both \( C1 \) and \( C2 \). The F value between \( C2 \) and \( C1 \) is defined as (Equation [12]):

\[
F(C2, C1) = \frac{2 \times R(C2, C1) \times P(C2, C1)}{R(C2, C1) + P(C2, C1)}
\]

(12)

where (Equations [13] and [14]),

\[
R(C2, C1) = \frac{n_{ij}}{n_c}
\]

(13)

and

\[
P(C2, C1) = \frac{n_{ji}}{n_t}
\]

(14)

The F value of the category \( C1 \), is the maximum F value obtained at any cluster in the set \( S2 \) (Equation [15]):

\[
F(C1) = \max_{i} F(C_i, C)
\]

(15)

Here, \( k \) is the number of the clusters in \( S_i \).

The next step is to assess whether the network node centrality values from the link connection network are not significantly different from those of the semantic connection network.

Toward this aim, a group of network node features is defined first.

The centrality includes in_degree, out_degree, betweenness, in_closeness, and out_closeness. They are indicators of importance of a node in a network. Each addresses the importance from a different perspective. For instance, the degree gauges the extent to which a node links to other nodes in a network; the betweenness measures the degree to which other nodes link one another by a node; and the closeness assesses the degree to which a node is far away from other nodes. The density is used to measure a network rather than a node. It measures the extent to which all nodes in a network are connected among themselves.

The degree of a node refers to the number of the connections maintained by that node in a network. The degree of a node indicates its influence or interaction strength in the network. The degree can be divided into out_degree and in_degree in a directional network.

For the link connection network, or the TTWILM, the in_degree of a node is defined (Equation [17]):

\[
\text{In\_degree}(i) = \frac{\sum_{j=1}^{n} w_{ij}}{n-1}
\]

(17)

The out_degree of a node is defined (Equation [18]):

\[
\text{Out\_degree}(j) = \frac{\sum_{i=1}^{n} w_{ij}}{n-1}
\]

(18)

For the adjusted semantic connection network, or the ASSM, the degree of a node is defined (Equation [19]):

\[
\text{Degree}(i) = \frac{\sum_{j=1}^{n} t_{ij}'}{n-1}
\]

(19)

The betweenness refers to the extent to which a node sits between other nodes in a network. It indicates the extent to which a node plays a role as an intermediary in the network. Definition of the betweenness for the link connection network (or the TTWILM) and the adjusted semantic connection network (or the ASSM) is shown in the following equation (Equation [20]).

\[
\text{Betweenness}(i) = \begin{cases} 
\sum_{j=1}^{n} \sum_{k=1}^{n} w_{jk} (i) / n^2 - 3n + 2, & \text{TTWILM} \\
\sum_{j=1}^{n} \sum_{k=1}^{n} t_{jk}' (i) / n^2 - 3n + 2, & \text{ASSM}
\end{cases}
\]

(20)

The closeness refers to the extent to which a node reaches to other nodes in a network. It is measured as the reverse of the average distance between the node and other nodes in the network. The distance between two nodes \( (d_{ij}) \) in a network is defined as the shortest path from node \( i \) to node \( j \).

For the link connection network, or the TTWILM, the in_closeness of a node is defined (Equation [21]):

\[
\text{In\_closeness}(i) = \frac{n-1}{\sum_{j=1}^{n} d_{ij}}
\]

(21)

The out_closeness of a node is defined (Equation [22]):

\[
\text{Out\_closeness}(j) = \frac{n-1}{\sum_{i=1}^{n} d_{ij}}
\]

(22)
For the adjusted semantic connection network, or the ASSM, the closeness of a node is defined (Equation [23]):

\[
\text{Closeness}(i) = \frac{n - 1}{\sum_{j=1}^{n} d_{ij}}
\]  

(23)

For the link connection network and the semantic connection network, the density of a network is defined in the following equation. Here, \(m\) is the sum of the weights of all the edges/ties in the network (Equations [24] and [25]).

\[
\text{Density} = \frac{m}{n \times (n-1)}
\]

(24)

\[
m = \begin{cases} 
\sum_{i=1}^{n} \sum_{j=1}^{n} t_{ij}, & \text{ASSM} \\
\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}, & \text{TTWILM}
\end{cases}
\]

(25)

A clan is a completed subnetwork where all nodes in the clan are connected to one another. Cluster is a broader concept. Nodes in a network can be clustered based on a variety of criteria. The clan method is one of many ways to cluster the nodes. A clan can be defined by a variety of ways. The clan concept was introduced because it was a potential clustering method for the node cluster comparison analysis between the two networks. In this study, the concept of the \(N\)-clan was used. In an \(N\)-clan, the distance between any two nodes \(i\) and \(j\) is smaller than or equal to \(N\). It is defined as follows (Equation [26]).

\[
d_{ij} \leq N,
\]

(26)

The relationships between the proposed research objectives and the research methods were summarized as follows. The comparison between the link connection network and the semantic connection network generated from the navigation guidance system was conducted using the SNA method, which was twofold. First is the clustering result comparison between the clusters from the link connection network and those from the semantic connection network. Toward this objective, the CONCOR method and the \(F\)-value cluster comparison analysis method were applied. The second is the network node feature comparison. In order to achieve this objective, the density and centrality (in_closeness, out_closeness, betweenness, in_degree, and out_degree) in the link connection network were compared with those in the semantic connection network using the \(t\)-test inferential analysis method. The correlation analyses between the centralities of a node and its information characteristics in the navigation guidance system were conducted using the Pearson correlation inferential analysis method. A topic node was recommended to optimize and enhance the topic-based navigation guidance system if the large centrality differences between the adjusted semantic connection network and the link connection network were detected. The recommended results were also evaluated by domain experts.

The SNA was conducted using the SNA package UCINET for Windows (Version 6.0), and the inferential statistical analysis was carried out by using SPSS (Version 22).

**Summary**

The relationships between the proposed research objectives and the research methods were summarized as follows. The comparison between the link connection network and the semantic connection network generated from the navigation guidance system was conducted using the SNA method, which was twofold. First is the clustering result comparison between the clusters from the link connection network and those from the semantic connection network. Toward this objective, the CONCOR method and the \(F\)-value cluster comparison analysis method were applied. The second is the network node feature comparison. In order to achieve this objective, the density and centrality (in_closeness, out_closeness, betweenness, in_degree, and out_degree) in the link connection network were compared with those in the semantic connection network using the \(t\)-test inferential analysis method. The correlation analyses between the centralities of a node and its information characteristics in the navigation guidance system were conducted using the Pearson correlation inferential analysis method. A topic node was recommended to optimize and enhance the topic-based navigation guidance system if the large centrality differences between the adjusted semantic connection network and the link connection network were detected. The recommended results were also evaluated by domain experts.

The SNA was conducted using the SNA package UCINET for Windows (Version 6.0), and the inferential statistical analysis was carried out by using SPSS (Version 22).

**Analysis and Results**

**Basic Information of the Navigation Guidance System**

All the topics located in the Health Topic entry page were specified and extracted, each topic was associated with a topic page and keywords from the topic page were extracted,
and the related topics of the topic and other information (general information, technical information, publications, and multimedia) were recorded. Data collection started in September 2013 and ended in January 2014. The data collection process took approximately four months.

The number of the topics in the navigation guidance system was 207. The number of the extracted raw keywords was 820. After the keyword validation process, the number of the left keywords was 624.

The identified topics served as the nodes, and the related link connections and semantic connections served as edges/ties to form the link connection network and the semantic connection network for comparison analysis, respectively. TTWILM describes the link connection network. The semantic connection network and the adjusted semantic connection network are described by the SSM and the ASSM, respectively. Because these networks were built based on the same topic set, the number of nodes in these networks was the same (207). The number of edges/ties in the link connection network was 441 whereas the number of edges/ties in the semantic connection network was 21,038. Because the cut-off point value was set to 0.0518 and the edges whose values below the cut-off point were removed, the final number of the edges in the adjusted semantic connection network was 6,197.

The average distances were 5.132, 1.013, and 1.722 in the link connection network, the semantic connection network, and the adjusted semantic connection network, respectively.

The visual display of the link connection network is illustrated in Figure 2. In Figure 2, the topic nodes situated on the left side of the display were isolated topic nodes. They were not connected to any other topic nodes in the navigation guidance system. The large size of the isolated topics suggested that it is necessary to enhance the navigation guidance system, and these topics are potential topic candidates for enhancement. The experts believed that some of the isolated topics should be connected to other topics. In fact, some of them were judged as related topics to other topics by the experts in this study.

**Clustering Comparison Analysis**

The CONCOR algorithm was applied to the link connection network or the TTWILM; it resulted in seven node blocks or clusters ($k = 7$). The number of the nodes in each of the seven clusters is shown in Table 1.

The $k$-means clustering algorithm was applied to the semantic connection network or the SSM ($k = 7$). The number of the nodes in each of the clusters is also shown in Table 1.
According to Equations (12)–(16), the similarity between the CONCOR clustering result and the $k$-mean clustering result was equal to 0.236. The similarities of the subclusters between the CONCOR clustering result and the $k$-mean clustering result are demonstrated in Table 1. The small $F$ value (0.236) and other small $F$ values of the subclusters indicate that there was a large difference between the link connection network and the semantic connection network.

A network density analysis was conducted in the link connection network and the semantic connection network, respectively. The results are illustrated in Table 2. The results show that the densities of the semantic connection network were much larger than the densities of the link connection network. The density of the semantic connection network (0.052) was much smaller than that of the link connection network (0.003). It means that the former had more edges/connections than the latter.

Notice that there were many isolated topics (34 topic nodes). They accounted for 16.42% (34 of 207) of the total nodes in the navigation guidance system. They formed an independent block in the CONCOR analysis. Cluster 7 (density = 0.000) of the link connection network in Table 2 confirms that the isolated topics or nodes in the link connection network were:

- Poisons; Tularaemia; Indigenous populations; Genetics; Medical waste; Mortality; Diphtheria; Innovation; Accident, Radiation; and Intestinal diseases, Parasitic.

The visual display of the clustered link connection network is exhibited in Figure 3. In Figure 3, the topic nodes with the same icon shape and color belong to the same cluster.

t-Test Results

A series of the $t$ tests were carried out to compare the link connection network and the semantic connection network from a different perspective. Because the comparison focused on the network node relationships, the adjusted semantic connection network was used. Test results are shown in Tables 3–5.

Because of the in_degree and out_degree in the directed link connection network, they were compared with the degree in the semantic connection network, respectively. For the same reason, the in_closeness and out_closeness in the link connection network were treated separately. The resultant $p$ values for the in_degree, out_degree, betweenness, in_closeness, and out_closeness were .000, .000, .006, .000, and .000, respectively. They were all smaller than the significance level .05. It suggests that there were significant differences between the two tested networks in terms of the in_degree, out_degree, betweenness, in_closeness, and out_closeness. These test results corroborate the previous clustering comparison results from a different perspective.

Because the testing results show the significant differences, it is necessary to analyze what caused the significant differences between the two networks at the topic level.

The top 20 topics with high scores in different categories of the in_degree, out_degree, betweenness, in_closeness, and out_closeness in the two networks are listed in Tables 6–8, respectively.

Reliability of Clusters

In the previously conducted clustering comparison analysis, the CONCOR method was applied to the link connection network and resulted in seven clusters. The $k$-means clustering method was applied to the semantic connection network and it yielded seven clusters. The $k$-means clustering method was applied to the link connection network; seven clusters were also produced. The result from the link connection network is exhibited in Figure 3. In Figure 3, the topic nodes with the same icon shape and color belong to the same cluster.

TABLE 1. Clustering analysis result summary.

| Clusters | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|----------|----|----|----|----|----|----|----|
| Size in the TTWILM | 51 | 23 | 43 | 40 | 9 | 7 | 34 |
| Size in the SSM | 26 | 101 | 18 | 11 | 23 | 5 | 23 |
| Number of the shared nodes | 13 | 18 | 12 | 1 | 2 | 0 | 2 |
| F value | 0.338 | 0.290 | 0.393 | 0.397 | 0.125 | 0.000 | 0.105 |
| Final $F$ value, $k$ = 7, number of the total nodes = 207 |

TABLE 2. Summary of the density analysis.

| Cluster | Density of the link connection network | Density of the semantic connection network |
|---------|----------------------------------------|-------------------------------------------|
| Block1/Cluster1 | .014 | .205 |
| Block 2/Cluster2 | .019 | .233 |
| Block 3/Cluster3 | .008 | .046 |
| Block 4/Cluster4 | .006 | .192 |
| Block 5/Cluster5 | .003 | .196 |
| Block 6/Cluster6 | .013 | .576 |
| Block 7/Cluster7 | .000 | .152 |
| Entire network | .003 | .052 |
network was compared with the result from the semantic connection network using the same clustering method (the k-means clustering method). The new similarity between the two resultant cluster sets was equal to .117. The summary of this test is demonstrated in Table 9. It is apparent that the two F values (one was equal to .117 and the other was equal to .236) of the two comparisons were quite low. It implies that the conclusion on the previous clustering comparison analysis is reliable.

Impact of Cut-Off Points

Next, let us discuss the impact of the different cut-off points on the t-tests results. The adjusted semantic connection network was derived from the semantic connection network after a cut-off point was selected and applied to the network. Because the selection of a cut-off point would affect the removal of the edges in the adjusted semantic connection network, it would have an impact on the ultimate t-test results. Equation (9) indicates that the average of all the weights in the SSM was selected as the cut-off point value. A series of the cut-off points were set and corresponding t tests
were conducted in order to ascertain the impact of the different cut-off points on the t-test results. The group of six additional cut-off points is defined in Equation (27). The $SD$ stands for standard deviation. It means that six additional t-tests were carried out.

$$CP(q) = Mean + q \times SD = 0.0518 + q \times 0.0053, \quad q = -3, -2, -1, 1, 2, 3.$$  

(27)

The new t-test results are shown in Tables 10 and 11, respectively. In Tables 10 and 11, the $p$ values of the in_degree, out_degree, in_closeness, and out_closeness in the six additional tests remained the same (.000). However, as the cut-off point value decreased, the corresponding $p$ value of the betweenness result decreased. But, in all the cases, the $p$ values were still smaller than the significance level .05. It is evident that the change of the cut-off points did not change the test conclusion. It echoes the previous t-test results.

### Clan Analysis

Clique/clan analysis can also lead to a group of node clusters in a network. The reason why the CONCOR method, rather than the clique/clan analysis, was utilized in this study is that some clusters may overlap with other clusters in the clique/clan analysis method. Our pilot study shows that the cluster overlapping rates in the clique/clan analysis method were very high. On the other hand, the clusters from the semantic connection network using the k-means clustering method were mutually exclusive. In other words, there were no overlapping parts between any two resultant clusters. It makes the sound and effective cluster comparison between the clustering results from the link connection network and the clustering results from the adjusted semantic connection network impossible.

If the overlapping rate is defined in Equation (28) and the clan analysis is applied to the link connection network, then the overlapping rates between two resultant clans can be calculated. In Equation (28), $IC_l$ refers to the number of items in the set $C$. $MIN(X, Y)$ is equal to $X$ if $X$ is smaller than $Y$. Otherwise, it is equal to $Y$.

$$Overlapping\_rate = \frac{|C_i \cap C_j|}{MIN(|C_i|, |C_j|)}$$  

(28)

| Table 5. t-test result for closeness. |
|-------------------------------------|
| The link connection network         |
| In_closeness | Closeness | Out_closeness | Closeness |
| Mean         | 0.008     | 0.586        | 0.006     | 0.586     |
| SD           | 0.005     | 0.056        | 0.001     | 0.056     |
| $p$ value $= 0.000$, $t$ value $= -146.483$, $df = 206$ | |

| Table 6. Top 20 topics with high degree scores in the two networks. |
|---------------------------------------------------------------|
| Rank | Topic | Out_degree | Topic | In_degree | Degree |
|------|-------|------------|-------|-----------|--------|
| 1    | Infant nutrition | .0058 | Nutrition | .0175 | Schistosomiasis | .1050 |
| 2    | Prisons | .0043 | Food safety | .0174 | Hepatitis | .1021 |
| 3    | Nutrition disorders | .0041 | Mental health | .0093 | Poliomyelitis | .1014 |
| 4    | Ashma | .0040 | Chronic diseases | .0078 | Leishmaniasis | .0943 |
| 5    | Blindness | .0039 | Tropical diseases | .0073 | Measles | .0914 |
| 6    | Food additives | .0039 | Child health | .0073 | Dengue | .0886 |
| 7    | Diet | .0037 | Health systems | .0068 | Filariasis | .0880 |
| 8    | Child development | .0037 | Infectious diseases | .0055 | Rubella | .0799 |
| 9    | Rubella | .0036 | Health workforce | .0053 | Rotavirus infections | .0783 |
| 10   | Violence | .0033 | HIV/AIDS | .0051 | Escherichia coli infections | .0770 |
| 11   | Health services | .0032 | Water | .0050 | Avian influenza | .0765 |
| 12   | Midwifery | .0032 | Zoonoses | .0050 | Blindness | .0750 |
| 13   | Condoms | .0031 | Injuries | .0035 | Mycobacterium ulcerans | .0744 |
| 14   | Drinking water | .0030 | Environmental health | .0031 | Rabies | .0741 |
| 15   | Breastfeeding | .0030 | Patient safety | .0031 | Trachoma | .0728 |
| 16   | Obesity | .0030 | Vaccines | .0030 | Urban health | .0725 |
| 17   | Melamine | .0027 | Technology, Health | .0029 | Campylobacter | .0715 |
| 18   | Mental disorders | .0026 | Immunization | .0028 | Listeria infections | .0707 |
| 19   | Food, genetically modified | .0025 | Obesity | .0027 | Onchocerciasis | .0673 |
| 20   | Mental health | .0024 | Pharmaceutical products | .0026 | Suicide | .0670 |
The summary of the overlapping rates are listed in Table 12. It was found that, among the 21 resultant overlapping rates, only 7 overlapping rates were equal to .000. The maximum overlapping rate among this pool reached .350, and average overlapping rate was .082.

**Correlation Between Centrality and Information Characteristics of a Topic**

Observe that, in the WHO portal, each topic web page includes four information categories (general information,
technical information, multimedia, and publication). Each category contains a group of the hyperlinks related to the topic. These categories provide additional detailed information on the topic. Nodes/topics with high scores of the in_closeness, out_closeness, betweenness, in_degree, and out_degree are pivotal nodes/topics in a network are central and critical to the network. They are supposed to contain rich and detailed information resources. In this situation, the topics with the high centrality scores in the network are supposed to correspond to a relatively large number of the links in each of the four categories. To examine this claim, a series of correlation analyses were conducted to ascertain the correlations between the network node centrality parameters of a topic (the in_closeness, out_closeness, betweenness, in_degree, and out_degree) and the number of the links in each of the categories (general information, technical information, multimedia, and publications) of the topic, respectively.

The Pearson correlation was applied for this purpose. Because there were four additional information categories and five categories of the network node centrality parameters, there were 20 correlation tests. The generated correlation test results are shown in Table 13. The results in Table 13 indicate that there were no strong correlations between the network node centrality characteristics of a topic and the number of the links in each of the four categories of the topic.

**Pivotal Nodes**

Additional information can be added to enrich the pivotal nodes identified in the study because these nodes are central to the navigation guidance system. These refer to the topics with high centrality scores in the link connection network. For example, if a node in the network has a high in_degree score, it means that many other topics cite this topic for a variety of reasons. The more a topic is cited, the more important it is. As a result, more information is needed to support the topic in

---

**TABLE 9. Summary of the new cluster comparison.**

| Clusters | C1 | C2 | C3 | C4 | C5 | C6 | C7 |
|----------|----|----|----|----|----|----|----|
| Size in the **TTWILM** | 5  | 4  | 6  | 12 | 172| 5  | 3  |
| Size in the **SSM** | 18 | 101| 26 | 23 | 23 | 11 | 5  |
| Number of the shared nodes | 5  | 2  | 1  | 1  | 22 | 0  | 0  |
| **F** value | .435| .038| .063| .057| .226| .000| .000|
| Final **F** value = 0.117, **k** = 7, number of the total nodes = 207 |

---

**TABLE 10. Summary of the descriptive statistic results of the additional t-tests.**

| Degree | Mean | .036| .037| .039| .042| .044| .046|
|--------|------|-----|-----|-----|-----|-----|-----|
| SD     | .019 | .020| .020| .021| .021| .021|
| Betweenness | Mean | .002| .002| .002| .002| .002| .002|
| SD     | .002 | .002| .002| .002| .002| .002|
| Closeness | Mean | .551| .562| .574| .601| .616| .636|
| SD     | .049 | .050| .053| .061| .066| .073|

---

**TABLE 11. Summary of the inferential statistic results of the additional t-tests.**

| In_degree |  |  |  |  |  |  |  |
|-----------|---|---|---|---|---|---|---|
| t value  | −26.389| −26.936| −27.350| −28.599| −29.533| −30.557|
| p value  | .000 | .000 | .000 | .000 | .000 | .000 |
| Out_degree |  |  |  |  |  |  |  |
| t value  | −26.386| −26.892| −27.297| −28.514| −29.432| −30.433|
| p value  | .000 | .000 | .000 | .000 | .000 | .000 |
| Betweenness |  |  |  |  |  |  |  |
| t value  | 2.437 | 2.544 | 2.650 | 2.909 | 3.051 | 3.199 |
| p value  | .016 | .012 | .009 | .004 | .003 | .002 |
| In_Closeness |  |  |  |  |  |  |  |
| t value  | −156.276| −157.219| −152.002| −139.690| −131.858| −123.006|
| p value  | .000 | .000 | .000 | .000 | .000 | .000 |
| Out_Closeness |  |  |  |  |  |  |  |
| t value  | −158.739| −159.569| −154.095| −141.289| −133.139| −123.929|
| p value  | .000 | .000 | .000 | .000 | .000 | .000 |
| df       | 206 |  |  |  |  |  |  |
the navigation system. The topic node *Mental disorder* had the high centrality scores \((\text{in\_degree} = .0013, \text{out\_degree} = .0026, \text{betweenness} = .0000, \text{in\_closeness} = .0169, \text{and\_out\_closeness} = .0064)\) in the link connection network. However, it contained a small number of the links in the four information categories (general information = 0, technical information = 0, publications = 0, and multimedia = 0). The average numbers of the links in the four categories (general information, technical information, publications, and multimedia) in the network are 1.391, 1.787, 1.589, and .778, respectively. It is apparent that the numbers of the links in the four categories for the topic node *Mental disorder* were lower than their corresponding average values. Consequently, the topic node *Mental disorder* was recommend to enrich its links in the four information categories (general information, technical information, publications, and multimedia).

### Visual Displays

The nodes whose corresponding centrality values were below the means were removed in the following displays. The removal was done in order to effectively visualize the degree distribution, betweenness distribution, and closeness distribution of the salient nodes in the adjusted semantic connection network.

Figure 4 gives the visual display of the degree in the adjusted semantic connection network. The mean of the degree scores in the network was .0406, and the number of the displayed nodes was 95. Figure 5 presents the visual display of the betweenness in the adjusted semantic connection network. The mean of the betweenness scores in the network was .0018, and the number of the displayed nodes was 72. Figure 6 illustrates the visual display of the closeness in the adjusted semantic connection network. The mean of the closeness scores in the network was .5859, and the number of the displayed nodes was 98.

In Figures 4–6, the size and color of a node icon correspond to its centrality value, and the thickness of an edge is commensurate with its connection strength. The edge weight cut-off values in Figures 4–6 were .200, .200, and .150, respectively. These figures were used in the subsequent discussion. The adjusted semantic connection network played an important role in the topic recommendation analysis. For instance, when a topic had high closeness score, high betweenness score, and high degree score, its related topics were considered as potential candidates for the topic recommendation. The visual displays of closeness, betweenness, and degree in the adjusted semantic connection network assisted people in identifying the potential candidates for the topic recommendation and understanding the topic recommendation analysis.

### Topic Recommendation and Evaluation

The procedure specifying the weak topic node candidates included the following: calculated the absolute node centrality differences between the adjusted semantic connection network and the link connection network in all the five centrality categories, respectively; formed five centrality difference lists, ranked each of the five lists in a descending order, respectively; and kept the top 50 topics in each of the ranked lists and other topics were removed. Finally, if a topic meets these conditions, it is considered a weak node. For

### Table 12. Summary of the overlapping rates for the resultant clans.

| Clan        | Common nodes | Minimum number | Overlapping rate |
|-------------|--------------|----------------|-----------------|
| Clan1 versus Clan2 | 0             | 21             | .000            |
| Clan1 versus Clan3 | 2             | 21             | .095            |
| Clan1 versus Clan4 | 1             | 20             | .050            |
| Clan1 versus Clan5 | 2             | 21             | .095            |
| Clan1 versus Clan6 | 1             | 20             | .050            |
| Clan1 versus Clan7 | 1             | 17             | .059            |
| Clan2 versus Clan3 | 0             | 21             | .000            |
| Clan2 versus Clan4 | 0             | 20             | .000            |
| Clan2 versus Clan5 | 6             | 23             | .261            |
| Clan2 versus Clan6 | 2             | 20             | .100            |
| Clan2 versus Clan7 | 2             | 17             | .118            |
| Clan3 versus Clan4 | 1             | 20             | .050            |
| Clan3 versus Clan5 | 2             | 21             | .095            |
| Clan3 versus Clan6 | 5             | 20             | .250            |
| Clan3 versus Clan7 | 1             | 17             | .059            |
| Clan4 versus Clan5 | 0             | 20             | .000            |
| Clan4 versus Clan6 | 2             | 20             | .100            |
| Clan4 versus Clan7 | 0             | 17             | .000            |
| Clan5 versus Clan6 | 7             | 20             | .350            |
| Clan5 versus Clan7 | 0             | 17             | .000            |
| Clan6 versus Clan7 | 0             | 17             | .000            |
| Average       | 1.667         | 19.524         | .082            |

### Table 13. Summary of the correlation test results.

|                  | General Information | Technical Information | Publications | Multimedia |
|------------------|---------------------|-----------------------|--------------|------------|
|                  | \(p\) value | Pearson correlation | \(p\) value | Pearson correlation | \(p\) value | Pearson correlation | \(p\) value | Pearson correlation |
| Out\_degree     | .103       | .114                 | .003        | .204        | .005       | .193                 | .020       | .160                 |
| In\_degree      | .000       | .370                 | .000        | .389        | .000       | .391                 | .000       | .481                 |
| Betweenness     | .000       | .317                 | .000        | .404        | .000       | .342                 | .000       | .379                 |
| In\_closeness   | .000       | .328                 | .000        | .380        | .000       | .337                 | .000       | .446                 |
| Out\_closeness  | .000       | .339                 | .000        | .391        | .000       | .331                 | .000       | .262                 |
FIG. 4. Display of degree in the adjusted semantic connection network. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

FIG. 5. Display of betweenness in the adjusted semantic connection network. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]
each weak node, its adjacent nodes in the adjusted semantic connection network were consulted. The adjacent nodes were recommended if their weights were larger than .400 in the adjusted semantic connection network and they were not within its adjacent node set in the link connection network. All the identified weak nodes and the recommended related topics are shown in Table 14. The number behind a related topic in Table 14 is the degree score between a weak topic and the related topic. Table 14 shows that 17 weak topics were identified and 41 related topics were recommended.

For instance, Congenital anomalies was a salient topic node in the semantic connection network, its degree score was .0629, its betweenness score was .0041, and its closeness score was .6417. Its in_degree and out_degree scores in the link connection network were .0003 and .0019, respectively, betweenness score in the link connection network was .0001, and in_closeness and out_closeness scores in the link connection network were .0049 and .0065, respectively. It is quite clear that there were significant differences between these parameters in these two networks. The implication is that the Congenital anomalies node in the link connection network needed to improve. Congenital anomalies, also known as birth defects, are structural or functional abnormalities, including metabolic disorders existing at birth and often before birth. It has strong semantic associations with other topics; unfortunately, it linked to only the topics Child health; Infant, Newborn; Nutrition; Disabilities; Rubella in the link connection network. This finding means that other

| Weak topics        | Recommended related topics                                      |
|--------------------|----------------------------------------------------------------|
| Schistosomiasis    | Leishmaniasis(0.4431); Filariasis(0.4012); Onchocerciasis(0.4006); Helminthiasis(0.4003) |
| Hepatitis          | Poliomyelitis(0.4000)                                          |
| Poliomyelitis      | Measles(0.5713); Mumps(0.5253); Smallpox(0.4690); Rubella(0.4387); Immunization(0.4239); Hepatitis(0.4000) |
| Leishmaniasis      | Onchocerciasis(0.4635); Schistosomiasis(0.4431)                |
| Measles            | Poliomyelitis(0.5713)                                          |
| Filariasis         | Onchocerciasis(0.4402); Mycobacterium ulcerans(0.4153); Schistosomiasis(0.4012) |
| Dengue             | Yellow fever(0.5677); Typhoid fever(0.5072)                    |
| Rubella            | Mumps(0.4467); Poliomyelitis(0.4387)                           |
| Rotavirus          | Campylobacter(0.4940); Listeria infections(0.4819)            |
| infections         | Escherichia coli infections(0.4225)                            |
| Mycobacterium ulcerans | Filariaisis(0.4153); Tropical diseases(0.4028)           |
| Rubies             | Chagas disease(0.4500); Anthrax(0.4273); Infectious disease(0.4151) |
| Listeria infections| Escherichia coli infections(0.5949); Rotavirus infections(0.4819); Campylobacter(0.4758); Salmonella(0.4104) |
| Onchocerciasis     | Leishmaniasis(0.4635); Filariaisis(0.4402); Schistosomiasis(0.4006) |
| Suicide            | Mental health(0.4072)                                          |
| Anaemia            | Nutrition disorders(0.4292)                                    |
| Congenital anomalies| Preterm birth(0.5219); Nutrition disorders(0.4289)          |
| Helminthiasis      | Schistosomiasis(0.4003)                                        |
topics as its related topics were missing in the navigation guidance system. The topics Preterm birth and Nutrition disorders, which did not appear in the link connection network, were found in the adjusted semantic connection network. In addition, the degree score between Congenital anomalies and Preterm birth in the adjusted semantic connection network was .5219 and the degree score between Congenital anomalies and Nutrition disorders in the adjusted semantic connection network was .4289. The two degree scores are larger than .4000. Thus, they were recommended as additional links to the topic in the navigation guidance system.

There is another example. Suicide was a weak topic node. In the semantic connection network, its degree score was .0670, its betweenness score was .0068, and its closeness score was .6890. Its in_degree score and out_degree score in the link connection network were .0002 and .0000, respectively, the betweenness score was .0000, and the in_closeness score and out_closeness score were .0049 and 0.0048, respectively. It is obvious that there were significant differences between these two networks in terms of these parameters. It only linked to the topic Depression in the guidance system. It suggests that the Suicide topic in the navigation guidance system needed to be enhanced. The topic Suicide has strong semantic associations with other topics, ranging from depression, to personality disorder, to schizophrenia, and to neurological disorders. It means that the topic Suicide should link to more related nodes in the navigation guidance system. The node Mental health meets the discussed conditions, and the degree score (.4072) between Mental health and Suicide in the adjusted semantic connection network was larger than .4000. Therefore, it is recommended as the related link topic for Suicide in the guidance system.

Expert Evaluation

Table 14 illustrates a group of the recommended related topics for the navigation guidance system. The recommended related topics were evaluated by the two experts. The agreement analysis between the two evaluators, and the consistency between the results from the study and the results from the experts, was analyzed. The resultant \( \chi \) of the kappa analysis was equal to .788. It indicates that there was a substantial agreement between the two evaluators according to the criteria presented by Viera and Garrett (2005). Next, a chi-square analysis was carried out to compare the integrated results of the two evaluators with the results of this study. The resultant \( \chi \), df, and \( p \) value were 2.050, 1, and .152, respectively. The chi-square results shows there were no significant differences between the results from the evaluators and the results from this study. In other words, the recommended related topics from this study were acceptable.

Implication of the Study

The clustering analysis results of this study can be used to enhance the navigation guidance system by building hierarchical relationships among the topics, optimizing their relationships, and enriching information for the pivotal topics. These would make the system more reasonable, accurate, and effective.

One enhancement could be the establishment of a topic-hierarchy—based navigation guidance system. The \( k \)-means clustering analysis of the semantic connection network resulted in a natural set of the seven category clusters, which can be used to build a topic hierarchy. The clustered topics can form a category, and all the categories can generate a subject-oriented hierarchy. The subject-oriented hierarchy would provide more guidance that is effective for users than a linear and alphabetically ordered topic list.

The mixed research method of this study can be applied to other similar web portals or digital libraries that have a navigation guidance system.

Conclusion

Public health portals are important information channels for health users to access information for their health decision making and other health information needs. Therefore, studies on the evaluation of find-aids in a public health portal have both theoretical and practical significances. In this study, the WHO portal was selected and its topic-oriented navigation guidance system was investigated.

After all the topics of the navigation guidance system in the WHO portal were identified and selected, the topics were kept, the topic-related link relationships were preserved, and all the keywords in the topic web pages were extracted and validated. There were 207 topics. The three networks (the link connection network, semantic connection network, and adjusted semantic connection network) were established based on the topics and their relationships in the guidance system. The link connection network describes the link relationships among the topics determined by the related topic links, whereas the semantic connection network and adjusted semantic connection network describe the semantic relationships among the topics defined by the topic web page contents. The CONCOR method was applied to the link connection network and the \( k \)-means clustering method was applied to the semantic connection network. Each clustering method resulted in a node cluster set that contained seven clusters. The resultant cluster set from the link connection network was compared with that from the semantic connection network. The similarity between the two cluster sets was calculated. In_closeness, out_closeness, betweenness, in_degree, and out_degree between the link connection network and adjusted semantic connection network were also compared by using an inferential statistical analysis method, respectively. The correlation analyses between the network centrality and topic information characteristics were investigated in detail.

The findings show that there were significant differences between the link connection network and the semantic connection network generated from the navigation guidance system. The similarity between the two cluster result sets...
from the two networks was equal to .236. There were significant differences between the link connection network and the semantic network in terms of the density, in_closeness, out_closeness, betweenness, in_degree, and out_degree as well. There were no strong correlations between the centralities of a node and its topic information characteristics (general information, technical information, publications, and multimedia). The suggestions for enhancing the navigation guidance system, which were evaluated and corroborated by domain experts, were recommended.

The 17 weak nodes in the link connection network were identified, the 41 corresponding related topics were specified and recommended to enhance the navigation guidance system, the recommended topics were evaluated by two evaluators, and the detailed discussion was included. The feedback from the evaluators revealed that the study was important for health consumers, and the design of a navigation guidance system of a public health portal should involve both information professionals and health experts.

Future Research

Future research directions include, but are not limited to, the user transaction log analysis on the navigation guidance system. If the user transaction log of the web portal could be accessed and users’ traversal activity data on the navigation guidance system could be utilized and analyzed, then user-oriented patterns on the guidance system could be investigated and revealed. The results are generated from the end users’ perspective. In this study, the semantic connection network based on the web page contents of the topics served as a baseline for the evaluation analysis. If the evaluation of the guidance system could be conducted from both the topic web page contents’ perspective and end users’ perspective, the findings would be more interesting and convincing. Another future research direction is that the same research method presented in this study be applied to other similar public health portals. The portal is dynamic and the contents of the portal change to reflect new events and health concerns. If analysis of terms and term relationships in the portal are periodically conducted, it can provide a temporal change picture of the health terminology.

Acknowledgment

We would like to thank Dr. Jinghao Chen for his contribution to data collection.

References

Barnes, J. (1954). Class and communities in a Norwegian island parish. Human Relations, 7(1), 39–58.
Benigeri, M., & Pluye, P. (2003). Shortcomings of health information on the Internet. Health Promotion International, 18(4), 381–386.
Borgatti, S., & Foster, P. (2003). The network paradigm in organizational research: A review and typology. Journal of Management, 29(6), 991–1013.
Bosche, P., & Segers, M. (2013). Transfer of training: Adding insight through social network analysis. Educational Research Review, 8(4), 37–47.
Burt, R. (2000). The network structure of social capital. In B.M. Staw, R. I. Sutton (Eds.), Research in Organizational Behavior, 22, 345–423.
Chambers, D., Wilson, P., Thompson, C., & Harden, M. (2012). Social network analysis in healthcare settings: A systematic scoping review. PLoS ONE, 7(8), 1–10.
Cline, R., & Haynes, K. (2001). Consumer health information seeking on the Internet: The state of the art. Health Education Research, 16(6), 671–692.
Coleman, J. (1986). Social theory, social research, and a theory of action. American Journal of Sociology, 91(6), 1399–1335.
Cook, K., & Whitmeyer, J. (1992). Two approaches to social structure: Exchange theory and network analysis. Annual Review of Sociology, 18, 109–127.
Evers, K. (2006). eHealth promotion: The use of the Internet for health promotion. American Journal of Health Promotion, 20(4), 1–6.
Fox, S., & Duggan, M. (2013). Health online 2013. Pew Research Center’s Internet & American Life Project. Retrieved from Retrieved from http://www.pewinternet.org/files/old-media/Files/Reports/PIP_HealthOnline.pdf
Grabowicz, P., Ramasco, J., Moro, E., Pujol, J., & Eguiluz, V. (2012). Social features of online networks: The strength of intermediary ties in online Social media. PLoS ONE, 7(1), 1–9.
Gray, N., Klein, J., Noyce, P., Sesselberg, T., & Cantrill, J. (2005). Health information-seeking behavior in adolescence: The place of the internet. Social Science & Medicine, 60(7), 1467–1478.
Griffiths, F., Cave, J., Boardman, F., Ren, J., Pawlikowska, T., Ball, R., . . . Cohen, A. (2012). Social networks: The future for health care delivery. Social Science & Medicine, 75(12), 2233–2241.
Guo, L. (2012). The application of social network analysis in agenda setting research: A methodological exploration. Journal of Broadcasting & Electronic Media, 56(4), 616–631.
Hambrick, M. (2012). Six degrees of information: Using social network analysis to explore the spread of information within sport social networks. International Journal of Sport Communication, 5(1), 16–34.
Hoppe, B., & Reinelt, C. (2010). Social network analysis and the evaluation of leadership networks. Leadership Quarterly, 21(4), 600–619.
Kempe, D., Kleinberg, J., & Tardos, E. (2003). Maximizing the spread of influence through a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 137–146). New York, NY, USA: ACM.
Knoke, D., & Yang, S. (2008). Social network analysis. Thousand Oaks, CA: Sage.
Koehly, L., Peterson, S., Watts, B., Kempf, K., Vernon, S., & Grize, E. (2003). A social network analysis of communication about hereditary nonpolyposis colorectal cancer genetic testing and family functioning. Cancer Epidemiology Biomarkers & Prevention, 12(4), 304–313.
Korthage, R.R. (1997). Information storage and retrieval. New York: Wiley Computer Publications.
Kwait, J., Valente, T., & Celentano, D. (2001). Interorganizational relationships among HIV/AIDS service organizations in Baltimore: A network analysis. Journal of Urban Health: Bulletin of the New York Academy of Medicine, 78(3), 468–487.
Larsen, B., & Aone, C. (1999). Fast and effective text mining using linear-time document clustering. Paper presented at the KDD, August 15–18, San Diego, California.
Luke, D., & Harris, J. (2007). Network analysis in public health: History, methods, and applications. Annual Review of Public Health, 28, 69–93.
Mertens, F., Saint-Charles, J., & Mergler, D. (2012). Social communication network analysis of the role of participatory research in the adoption of new fish consumption behaviors. Social Science & Medicine, 75(4), 643–650.
Park, H. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. Connections, 25(1), 49–61.
Pow, J., Gayen, K., Elliott, L., & Raeside, R. (2012). Understanding complex interactions using social network analysis. Journal of Clinical Nursing, 21(19), 2772–2779.
