Concerns about health issues cover a wide spectrum. Consumer health information, which has become more available on the Internet, plays an extremely important role in addressing these concerns. A subject directory as an information organization and browsing mechanism is widely used in consumer health-related Websites. In this study we employed the information visualization technique Self-Organizing Map (SOM) in combination with a new U-matrix algorithm to analyze health subject clusters through a Web transaction log. An experimental study was conducted to test the proposed methods. The findings show that the clusters identified from the same cells based on path-length-1 outperformed both the clusters from the adjacent cells based on path-length-1 and the clusters from the same cells based on path-length-2 in the visual SOM display. The U-matrix method successfully distinguished the irrelevant subjects situated in the adjacent cells with different colors in the SOM display. The findings of this study lead to a better understanding of the health-related subject relationship from the users’ traversal perspective.

Introduction

Health information is important to the public. With the dramatic development of the Internet, more and more health information is available online for consumers. According to an investigation by Dickerson et al. (2004), 68% of patients who accessed the Web used it to search for health information. Due to the complicated nature of health information, its organization and retrieval is a long-standing research topic.

One important means for information organization is the subject directory, which arranges and divides objects into several subject groups based on their attributes and characteristics. A subject directory plays such an important role in information organization that it has been the topic of many studies. To deal with the huge amount of information on the Web, Website designers and researchers have been seeking to organize information in Web directories, because people, places, concepts, events, properties, and attributes can be classified or categorized according to some taxonomic scheme. Usually, a taxonomy is a subject-oriented system where the sibling subjects are exclusive. Taxonomies are often designed in a hierarchical structure and provide a holistic framework for users to browse information intuitively and conveniently.

The facet method exemplifies the inductive scientific method due to the synthesis of conceptual knowledge (Richmond, 1976). Facets can effectively handle either single or multiple values. In a faceted information system, selecting a term is equivalent to carrying out a disjunction operation over all the terms beneath the selected one (Quintarelli, Resmini, & Rosati, 2007). Designers have tried to make Web directories user-oriented to confront the challenges from both diverse users and from the dynamic information on the Internet. As a result, Web subject definitions and subject directory structures are usually loosely defined based on users’ preferences and usage. For this reason consumer health information Websites such as WebMD.com, Healthline.com, WomensHealth.gov, etc., utilize Web directories to organize online health information.
Since subject directories are widely used for Internet information, the evaluation of these Web directories becomes an urgent and vital task. Transaction log analysis, an effective user behavior analysis method, can be used for Web directory evaluation. A Web transaction log is automatically created and kept by a server, and it faithfully records the history of all requests from users. Because every traversal activity of users on a Web directory is recorded, these Web transaction logs provide truthful, complete, first-hand, and valuable sources for the evaluation of Web directories. The traversal relationships among nodes in a Web directory, which are fundamental and critical for Web directory usage analysis, are multiple and therefore high-dimensional. Information visualization techniques can transform such high-dimensional and abstract data or information into two-dimensional or three-dimensional presentations and effectively demonstrate sophisticated connections among objects in a visual display (Zhang, 2008). These techniques can be applied to the display of retrieved results and relationships among the retrieved items (Zhang & Korfhage, 1999; Zhang, 2001). The Self-Organizing Map (SOM) is an exploratory analysis method that visualizes the high-dimensional input data in a low-dimensional space and preserves the topology of the input data. The application of an information visualization method such as SOM provides a unique avenue for Web directory analysis.

Users’ traversal activities in a Website can reveal rich information about their information-seeking behavior in the Website. Users’ traversal activities in the context of a subject directory would definitely exhibit the relationships among the subject nodes in the directory. It is evident that users’ traversal lengths in a subject directory vary in different traversal activities. As the traversal length increases, the relevance between two subject nodes that are located at the two ends of the traversal path will typically decrease. This suggests that the subject clusters identified in the SOM analysis method based on a shorter path length may be more relevant than these identified in the SOM analysis method based on a longer path length. In other words, the SOM displays based on different path lengths can generate different subject cluster analysis results, and these subject clusters can be ranked based on different path lengths in terms of relevance. As a result, the subject cluster analysis results based on different path lengths can be used not only to identify terms related to a subject, but also to rank them as related terms in a subject directory. Related Web pages are usually listed under a subject node in a subject directory and a Web page may correspond to multiple subjects. If a group of Web pages are relevant to a subject, then they can be categorized under that subject in a subject directory and ranked against the other associated subjects that have been ranked based on different traversal lengths. Ranking both related subjects and related Web pages in a subject directory based on users’ traversal behavior would make the subject directory more user-centered.

It is widely recognized that information system design should be user-oriented rather than system-oriented, and Web-based subject directory design is no exception. Web-based subject directories are implemented in a variety of Web portals as an effective browsing mechanism or navigation tool. Traditional design principles, however, may no longer be suitable for the design of dynamic, user-oriented, and Web-based subject directories. For instance, traditional subject directory design principles are insufficient to answer such questions as: What constitutes a user-friendly subject directory, which subject category in a user-oriented directory should be broken into several subject subcategories, which subject categories should be merged into a new category, which subject categories should be related, and to what degree should the subjects be related? The academic or practical significance of the path analysis method rests on the fact that it provides people with first-hand data and a scientific method for decision-making about user-oriented subject directory design.

The purpose of this study is to examine (1) whether the proposed approach of integrating the U-matrix with SOM can significantly improve the performance of the traditional SOM method; (2) whether the subject cluster analysis results from the same cells based on path length 1 are better than the subject cluster analysis results from the adjacent cells based on path length 1 in the SOM display; and (3) whether the subject cluster analysis results from the same cells based on path length 1 are better than the subject cluster analysis results from the same cells based on path length 2 in the SOM display. A cell in the SOM display refers to the smallest organizational unit in the SOM grid. Adjacent cells of a cell refer to all the cells that are directly connected to the cell in the SOM grid. In other words, the study investigates the impact of path length in a subject directory on the association strength between the two subjects connected by the path, as well as the effectiveness of the proposed approach of integrating the U-matrix with SOM. The findings can be used to understand health-related subject relationships from the users’ traversal perspective, and to specify user-relevant health subject clusters.

Related Research

Health Informatics

Health informatics is a multidisciplinary research field, involving information science, computer science, medical science, and healthcare. It deals with the theory and practice of the acquisition, organization, retrieval, dissemination, and utilization of health information. Many studies have focused on the topic of semantic health information analysis. To identify related medical terms and clusters for the construction and revision of thesauri and classification systems, Zhang, Wolfram, Wang, Hong, and Gillis (2008) used a multidimensional-scaling approach to analyze frequently used medical-topic terms in queries submitted to a Web-based consumer health information system. Zhang and Wolfram (2009) studied obesity, a more specific health topic, from a public health portal transaction log and identified the relationships between obesity-related focus keywords.
and their co-occurring keywords. Ong, Chen, Sung, and Zhu (2005) presented the Newsmap, which generated a hierarchical knowledge map for online Chinese health news. It employed an alphabetical hierarchical structure and a visual island display which allowed users to browse the broader categories at a high level and the narrower subcategories at a low level.

**Self-Organizing Map**

SOM is one of the most prominent artificial neural network models, introduced by Kohonen (1982). In this model, objects in a high-dimensional signal space are trained by an unsupervised learning algorithm and then visualized in a low-dimensional feature map (Kohonen, 1997). One of the prominent attributes of this feature map is that it preserves the topological properties of the input objects. Similar objects are mapped onto adjacent grids (Ding & Patra, 2007).

Semantic relationships among objects like documents are usually multiple. The multiple relationships can be effectively described in an object–object proximity matrix if the relationship between two objects is clearly defined. The proximity matrix corresponds to a space. The dimensionality of the proximity matrix depends on the number of the described objects in the matrix. The dimensionality can reach to hundreds or even greater, but the semantic relationships preserved in the high-dimensional matrix cannot be displayed and perceived by people in the hyperspace if the dimensionality is greater than 3. Therefore, the preserved relationships in the hyperspace are not easily understood by users. If the dimensionality of the space is reduced to three or two by a conversion method such as the SOM method or the Multi-dimensional Scaling method (Young & Householder, 1941), the objects can be displayed and the relationships can still be kept in the low and visible space. As a result, users can view the relationships of the objects in a simple and intuitive way.

The SOM method has been widely used in many fields, such as visualization of machine states, fault identification (Kohonen, Oja, Simula, Visa, & Kangas, 1996), feature extraction, computer vision (Kohonen, 1996), exploratory data analysis, and knowledge discovery in databases (Kohonen, 1999). It has been reported that the research on SOM application has increased by about 15% each year (Kohonen et al., 1996).

Chen, Fan, Chau, and Zeng (2003) focused on a specific medical domain and developed a meta-search tool named Cancer Spider, which clustered Web pages into regions for different topics on a two-dimensional SOM display. AMedPort (Chung, Bonillas, Lai, Xi, & Chen, 2006) also utilized the SOM technique to categorize and search medical Web pages in Arabic. A language-independent approach was also implemented to build a collection for searching and browsing non-English medical information.

The principles and basic algorithms of SOM have been addressed in an abundant literature (Kohonen, 1990, 1993a, 1993b). In general, there are two main learning algorithms, i.e., the sequential (also online, stochastic, or incremental) algorithm and the batch algorithm. In the latter algorithm, the weight vectors associated with the cells in the feature map are updated after all the input vectors are processed, while in the former algorithm the weight vectors are revised shortly after each input vector is processed. The batch algorithm is much faster than the sequential algorithm. A thorough comparison of the two algorithms was conducted by Fort, Letremy, and Cottrell (2001, 2002). Their findings show that the batch algorithm is superior to the sequential algorithm in computational simplicity, efficient training time, less distortion, omission of the tuning adaptation parameter, and stability of results. Ding and Patra (2007) have also argued that the batch algorithm does not suffer from the convergence problem caused by a changeable learning rate \( \alpha \). However, the batch algorithm yields worse object organization and visualization results and less balanced categories than the sequential algorithm. In addition, the batch algorithm results are heavily dependent on the weight vector initialization.

As we know, the differences between the weight vectors associated with the SOM cells cannot be directly displayed in the SOM feature map because of the weight vectors’ high-dimensional nature. Thus, the concept Unified Distance Matrix (U-matrix) was introduced by Ultsch and Siemon (1990) to attack the problem. The U-matrix can effectively reveal the differences among the weight vectors in the feature map. The Euclidean distances between the weight vector of a cell and the weight vectors of adjacent cells in the feature map are computed and assigned to the corresponding element of the U-matrix. As a result, the elements with high values represent cluster boundaries, while those with low values represent clusters themselves in the U-matrix. The values in the U-matrix are then converted to various colors to illustrate the differences. The length and width of the original U-matrix are \( 2n - 1 \), where \( n \) is the length of the SOM grid. The units in the intersection of even rows and odd columns in the U-matrix represent the distances between vertical neighboring cells in the SOM grid, while those in the intersection of even rows and even columns in the U-matrix stand for the distances between the diagonal neighboring cells in the SOM grid.

Ultsch (1992) presented a revised version of the U-matrix. The new U-matrix has the same length and width as the original SOM grid. The value in each unit of the new U-matrix is equal to the sum of the distances of a cell to all its immediate neighboring cells normalized by the largest occurring value in the SOM grid.

**Transaction Log Analysis**

Recently, the Web transaction log analysis method has attracted a lot of interest from researchers in various fields. Findings from Web transaction log analysis can be used for optimizing the topology structure of Websites, increasing Web traffic, enhancing online services, monitoring and preventing illegal and malicious online activities, etc.

In general, transaction log analysis can be divided into two fields, the analysis of queries submitted to a server, and
traversal path analysis of browsing activities. The analysis of submitted queries usually examines the query terms that users enter. For example, Ohura, Takahashi, Pramudiono, and Kitsuregawa (2002) proposed a query expansion method based on clustering users’ query terms. Shi and Yang (2007) applied the query transformation patterns to the generation of a list of related and ranked queries for users to revise and expand an initial input query. To better understand session characteristics, Wolfram, Wang, and Zhang (2007) modeled session characteristics by comparing a health transaction log with two different transaction logs.

Subject Directory Research

Browsing and query searching are two paradigms of seeking information. The former depends heavily on a well-organized and presented information structures such as subject directories and classifications. The role of classification in information organization has often been underestimated due to the misconception that users prefer query searching to browsing. In a study, Koch, Golub, and Ardö (2006) explored the transaction log of Renardus, a DDC-based Web service that provided both browsing and query searching features. Surprisingly, browsing activities were found to be clearly dominant, for two reasons. One was that most browsed pages were indexed by search engines, so that users started their Renardus navigation at a browsing page. The other was due to the interface layout, which encouraged browsing by highlighting the browsing structure on the top of the search box. The authors therefore claimed that browsing might be useful and dominate navigation in circumstances similar to Renardus.

In the Web space, subject directories are an efficient means for organizing information. Many studies have been done on how subject directories are automatically created and maintained in the Web environment. Yang and Lee (2004) applied a text-mining technique based on SOM to automatically create Web directories and organize Web pages into hierarchies. Later, Yang and Lee (2006) came up with NaviSOM in support of a navigational structure for WWW information.

To organize a huge amount of information with a taxonomy scheme without getting lost in its complicated structure, Börner, Hardy, Herr, Holloway, and Paley (2007) introduced a taxonomy visualization and validation (TV) tool to assist the semiautomatic validation and optimization of organizational schemes. The TV tool facilitated identifying and reclassifying both misclassified information entities and large classes, evaluating the homogeneity of existing classes, and examining the “well-formedness” of an organized schema.

In summary, while some studies have been done on health information semantic analysis, such as identifying related medical terms and clusters for the construction and revision of thesauri and classification systems, few studies have been conducted on the traversal path analysis in a health subject directory. Research on SOM has primarily focused on algorithm design and optimization. Although the SOM techniques have been widely applied to health and medical fields, studies on the SOM technique applications in health transaction logs, especially in health directory traversal, are scant in the literature. In the field of Web subject directory research, most research has elaborated on the issues of automatically creating and maintaining Web directories, while few studies have been done on the use of these Web subject directories.

In this study we investigate the transaction log of a customer health portal, HealthLink (http://healthlink.mcw.edu/), analyze the traversal paths in a health subject directory, produce a cluster analysis for the directory subjects by using the SOM techniques, and use a new U-matrix method to illustrate sophisticated adjacent cell relationships and to distinguish irrelevant clusters located in the adjacent cells in the SOM display.

Research Method Description

The HealthLink Subject Directory Description

The data source in this study comes from the Medical College of Wisconsin’s HealthLink. HealthLink aims to provide reliable medical information services to their patients and community. The HealthLink portal incorporates a health subject directory that provides health consumers with an information browsing mechanism in addition to query search. The subject directory is a three-level hierarchy. Its root, called the topic, serves as a start point for browsing. All nodes on the directory branches are called subjects. In the directory, there are 47 subjects/nodes, such as Allergies/Asthma, Alternative Medicine, and so on. Table 1 shows a full list of the subjects. All related articles are mapped onto the corresponding leaf node of the subject directory. Users can browse its

<table>
<thead>
<tr>
<th>Label</th>
<th>Subject</th>
<th>Label</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Aging</td>
<td>25</td>
<td>Infections</td>
</tr>
<tr>
<td>2</td>
<td>Allergies</td>
<td>26</td>
<td>Kidney disease</td>
</tr>
<tr>
<td>3</td>
<td>Alternative medicine</td>
<td>27</td>
<td>Liver</td>
</tr>
<tr>
<td>4</td>
<td>Arthritis</td>
<td>28</td>
<td>Men’s health</td>
</tr>
<tr>
<td>5</td>
<td>Back problems</td>
<td>29</td>
<td>Mental health</td>
</tr>
<tr>
<td>6</td>
<td>Brain nervous system</td>
<td>30</td>
<td>Musculoskeletal</td>
</tr>
<tr>
<td>7</td>
<td>Cancer</td>
<td>31</td>
<td>Neurological disorders</td>
</tr>
<tr>
<td>8</td>
<td>Children’s health</td>
<td>32</td>
<td>Nutrition and herbs</td>
</tr>
<tr>
<td>9</td>
<td>Cholesterol</td>
<td>33</td>
<td>Occupational health</td>
</tr>
<tr>
<td>10</td>
<td>Clinical trials</td>
<td>34</td>
<td>Organ transplants</td>
</tr>
<tr>
<td>11</td>
<td>Diabetes</td>
<td>35</td>
<td>Pain</td>
</tr>
<tr>
<td>12</td>
<td>Digestive disease</td>
<td>36</td>
<td>Physical medicine</td>
</tr>
<tr>
<td>13</td>
<td>Drugs medications</td>
<td>37</td>
<td>Preventive medicine</td>
</tr>
<tr>
<td>14</td>
<td>Emergency medicine</td>
<td>38</td>
<td>Public health</td>
</tr>
<tr>
<td>15</td>
<td>Endocrine system</td>
<td>39</td>
<td>Respiratory</td>
</tr>
<tr>
<td>16</td>
<td>Environmental health</td>
<td>40</td>
<td>Safety</td>
</tr>
<tr>
<td>17</td>
<td>Eye care</td>
<td>41</td>
<td>Skin diseases</td>
</tr>
<tr>
<td>18</td>
<td>Feet</td>
<td>42</td>
<td>Sports medicine</td>
</tr>
<tr>
<td>19</td>
<td>Genetics</td>
<td>43</td>
<td>Travel medicine</td>
</tr>
<tr>
<td>20</td>
<td>Hearing disorders</td>
<td>44</td>
<td>Vitamins</td>
</tr>
<tr>
<td>21</td>
<td>Heart disease</td>
<td>45</td>
<td>Weight control</td>
</tr>
<tr>
<td>22</td>
<td>High blood pressure</td>
<td>46</td>
<td>Wellness lifestyle</td>
</tr>
<tr>
<td>23</td>
<td>Immune disorders</td>
<td>47</td>
<td>Women’s health</td>
</tr>
<tr>
<td>24</td>
<td>Immunization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
collection by starting with the topic link named “Browse by Topic” (http://healthlink.mcw.edu/topics/) on the portal. The topic link directs users to its children nodes (47 subjects). Under each subject there is a list of relevant articles.

Notice that in this hierarchy structure, users can jump from the topic root to any subject nodes, from a subject node to its related articles, and from an article to related articles or related subject nodes. However, users cannot jump from a subject node to other subject nodes directly.

Web Log Description and Cleaning

A Web transaction log contains all the information about a client’s requests. Transaction log formats vary in different servers. The most commonly used formats are known as Common Log Format and Combined Log Format (Apache, 2008).

Web transaction log cleaning is a complicated and time-consuming task. In this study we focused on the traversal paths, such as the paths from the topic to a subject, from a subject to an article, from an article to another article, from an article to a subject, from a subject to the topic, and from an article to the topic. In order to identify all paths from a transaction log, all referrers and topic destination pages, the 47 subjects and the connected articles had to be identified and kept. Other data that made no contribution to traversal paths analysis were excluded from the log.

The Web log analysis software Web Log Explorer was used to filter out irrelevant data. The records whose method was “GET,” response code was “200,” and whose “referrer” and “destination page” were either pages like http://healthlink.mcw.edu/ (subject) or http://healthlink.mcw.edu/article/* or http://healthlink.mcw.edu/topic were considered. The data whose referrer was http://healthlink.mcw.edu/search/*, and whose destination page was http://healthlink.mcw.edu/content/* were excluded. In addition, the time range and other related parameters were also included.

Definitions of the SOM Input Matrices

In this study, two kinds of SOM visual configurations for the involved subjects and topic were generated. One was produced based on the paths whose traversal path lengths were equal to 1, and the other was based on the paths whose traversal path lengths were equal to 2. The reason for creating an SOM visual configuration based on traversal path length 2 is that in the investigated HealthLink health subject directory there was not a direct link between two different subject nodes. In other words, users could not jump from a subject to another subject in the directory. A visual SOM configuration for traversal path length 2 can also reveal the semantic connections among the subjects in the directory, as a visual SOM configuration for traversal path length 1 does. When the traversal path length is equal to 2, it involves three Web nodes: a referrer node, an intermediate node, and a destination node. The intermediate nodes can vary, appearing as a site map link, an article, or a subject node, while both the referrer nodes and destination nodes must be the topic, or subjects. The semantic relationships identified in the visual SOM configuration based on path length 2 would definitely provide additional information useful for subject relevance analysis.

The SOM input format requires an $m \times n$ traversal matrix. The rows ($m$) of the matrix represent the objects visualized in the SOM display space, and the columns ($n$) of the matrix define the attributes of the objects.

The method of constructing the matrix for traversal path length 1 is as follows. Assume a matrix $MP_1$, see Equation (1)) with $m$ rows and $n$ columns. Rank all the investigated Web pages alphabetically according to their URLs and number them from 1 to $m$. In Equation (1) $c_{ij}$ $(i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)$ stands for a cell of the matrix. Element $c_{ij}$ in the matrix for traversal path length 1 is defined as the number of the times that users traverse from Web page $i$ to Web page $j$. If there is no traversal occurrence from Web page $i$ to Web page $j$, then $c_{ij}$ is equal to 0.

$$MP_1 = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{pmatrix} \quad (1)$$

The matrix $MP_2$, see Equation (2)) is for traversal path length 2, where element $d_{ij}$ in the matrix for traversal path length 2 is defined as the number of times that users traverse from Web page $i$ via an intermediate node to Web page $j$.

$$MP_2 = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mn} \end{pmatrix} \quad (2)$$

Definition of the Proposed U-matrix

As described in the section Health Informatics, the second definition of the U-matrix had the same length and width as the original SOM grid. According to the definition of the U-matrix, the value of a cell in the U-matrix is determined by the distances between the cell and its immediate neighbor cells. Since the number of immediate neighbors for a cell which is located on the border or at the corner of the SOM display is usually smaller than that of a cell located in the central area of the SOM display, the sum of the distances between the cell on the border (or at the corner) and its immediate neighbors is normally smaller than that of the distances between a central cell and its immediate neighbors. Unfortunately, the circumstances are not differentiated in the definition of the second U-matrix, nor is the impact of the corner neighbors of a cell considered in the definition. The corner neighbors of a cell in the SOM grid are the cells whose corners are connected to a corner of the cell in the SOM grid. For instance, a typical cell $(s_{ij})$ in the middle
To solve these problems, we proposed a new definition of the map has four immediate neighbors (denoted by \( s_{ij-1} \), \( s_{i-1j} \), \( s_{i+1j} \), and \( s_{i+1j-1} \)) and four corner neighbors (denoted by \( s_{i-1j-1} \), \( s_{i-1j+1} \), \( s_{i+1j+1} \), and \( s_{i+1j+1} \)) (Figure 1). These corner neighbors of a cell, like its immediate neighbors, also affect the cell, though the impact of these corner neighbors may not be as strong as that of the immediate neighbors.

To solve these problems, we proposed a new definition of the U-matrix described as follows and implemented it in MatLab (MathWorks, Natick, MA), where SOM Toolbox was integrated.

Assume a rectangular SOM display \( S \) with \( m \) rows and \( n \) columns in Equation (3):

\[
S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \cdots & s_{mn} \end{pmatrix}
\]  

(3)

where \( s_{ij} \) \((i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)\) represents the weight vector associated with the corresponding cell.

The new U-matrix with \( m \) rows and \( n \) columns is defined in Equation (4):

\[
U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ u_{21} & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1} & u_{m2} & \cdots & u_{mn} \end{pmatrix}
\]  

(4)

where each element is represented by \( u_{ij} \) \((i = 1, 2, \ldots, m; j = 1, 2, \ldots, n)\).

The cell \( u_{ij} \) indicates the difference between a cell and both its immediate neighbors and corner neighbors. The distance between a cell and its immediate neighbors is defined as:

\[
D_1 = d(s_{ij-1}, s_{ij}) + d(s_{i-1j}, s_{ij}) + d(s_{i+1j}, s_{ij}) + d(s_{i+1j-1}, s_{ij})
\]  

(5)

where \( d(s_i, s_j) \) stands for the Euclidean distance between \( s_i \) and \( s_j \).

The distance between a cell and its corner neighbors is defined as:

\[
D_2 = d(s_{i-1j-1}, s_{ij}) + d(s_{i-1j+1}, s_{ij}) + d(s_{i+1j+1}, s_{ij}) + d(s_{i+1j-1}, s_{ij})
\]  

(6)

Because the impact of the immediate neighbors on the cell is stronger than the impact of the corner neighbors, the impact of the corner neighbors should be softened to some degree. Therefore, we have Equation (7):

\[
u_{ij} = \frac{D_1 + D_2}{\sqrt{2}}
\]  

(7)

Suppose \( s_{ip} \) and \( s_{jr} \) are two cells in the SOM grid, \( s_{ip} = (x_{ip,1}, x_{ip,2}, \ldots, x_{ip,n}) \) and \( s_{jr} = (x_{jr,1}, x_{jr,2}, \ldots, x_{jr,n}) \) are their weight vectors, respectively, where \( k \) is the dimensionality of the weight vectors. Then, their distance is defined as:

\[
d(s_{ip}, s_{jr}) = \sqrt{\sum_{w=1}^{k}(x_{ip,w} - x_{jr,w})^2}
\]  

(8)

Equation (7) suggests that the value of an element in the U-matrix is the average distance between the cell and its immediate and corner neighbors. In this equation, the impact of immediate neighbors and the impact of corner neighbors are considered differently. The reason why the distances to the corner neighbors are divided by \( \sqrt{2} \) is that the corner neighbors are situated at the diagonal position of \( s_{ij} \) (Figure 1). If the geometric distance between \( s_{ij} \) and \( s_{i-1j} \), and that between \( s_{i-1j} \) and \( s_{i-1j-1} \) is 1, then the geometric distance between \( s_{ij} \) and \( s_{i-1j-1} \) is \( \sqrt{2} \). Similarly, the geometric distance between \( s_{ij} \) and \( s_{i+1j+1} \) (\( s_{i+1j+1} \) or \( s_{i+1j+1} \)) is also \( \sqrt{2} \). The shorter the distance between two cells in the SOM grid, the stronger the impact, and vice versa. That is why the distances to the corner neighbors are divided by \( \sqrt{2} \).

For the cells on the borders or at the corners in the SOM display, only the distances to their legitimate neighbors are considered. For instance, if a cell is located on the top border of the map (but not at its corners), its weight vector is denoted by \( s_{ij} \) \((i = 1, j \neq 1, j \neq n)\). It has only three immediate neighbors (whose weight vectors are denoted by \( s_{ij-1}, s_{ij+1}, \) and \( s_{i+1j} \)) and two corner neighbors (whose weight vectors are denoted by \( s_{i+1j-1}, \) and \( s_{i+1j+1} \)). The final value of the cell in the U-matrix is calculated in Equation (9):

\[
u_{ij} = \frac{d(s_{ij-1}, s_{ij}) + d(s_{i-1j}, s_{ij}) + d(s_{i+1j}, s_{ij}) + d(s_{i+1j-1}, s_{ij})}{5}
\]  

(9)

The denominator in Equation (9) is 5 because there are only 5 neighbors, i.e., three immediate neighbors and two corner neighbors. For the cells on the bottom, left, or right border of the map, the corresponding U-matrix values are calculated by a similar method.

Now let us consider the corner cases. A cell located at the left top corner, whose weight vector is denoted by \( s_{11} \), has only two immediate neighbors (whose weight vectors are denoted by \( s_{21} \) and \( s_{12} \)) and one corner neighbor, \( s_{22} \). The value of the cell in the U-matrix is calculated in Equation (10):

\[
u_{11} = \frac{d(s_{12}, s_{11}) + d(s_{21}, s_{11}) + d(s_{21}, s_{12})}{3}
\]  

(10)

The denominator in Equation (10) is 3 because there are only 3 neighbors, i.e., two immediate neighbors and one
corner neighbor. For the cells at the right top (or the left bottom or the right bottom) corner, the corresponding U-matrix values are calculated in a similar way.

In summary, no matter where a cell is located, when calculating its U-matrix value, (1) only the distances to its legitimate neighbors are considered; (2) the distances to the corner neighbors are divided by \( \sqrt{2} \); and (3) the sums of all the distances are normalized by the total number of its legitimate neighbors. Equation (11) shows the common equation.

\[
U_{ij} = \frac{\sum d(s_{ij}, \text{immediate neighbors}) + \sum d(s_{ij}, \text{corner neighbors})}{\text{The total number of legitimate neighbors}}
\]

The high values in the U-matrix signify borders between clusters, and the low values indicate a cluster. The values of the U-matrix are used to produce the background colors for the SOM display. The background colors help people understand the object distribution in the SOM display.

**Analysis Method**

A subject directory is a hierarchical structure where subjects are located and semantic relationships among the subjects are presented. A subject directory can show sibling relationships, broad relationships, narrow relationships, and related relationships of a subject. After Web pages are mapped onto a Web subject directory, the semantic relationships among the Web pages are also illustrated. A subject directory can guide or direct users to relevant subject(s) and relevant Web pages during navigation. User traversal activities on a subject directory may reveal user-oriented semantic relationships among the subjects. This implies that if user traversal information on a subject directory is collected and the SOM approach is applied to the information, user-oriented semantic relations can be demonstrated and visually analyzed in the SOM contexts.

Since an SOM display preserves the topology of input subjects, the subjects projected onto the same cell or neighboring cells in the SOM display are considered to share similar characteristics. The subjects sharing similar characteristics can be clustered together. Subject clustering analysis in the SOM display is conducted at two levels. At the first level, the subjects that are projected onto the same cell are considered a cluster. At the second level, the subjects that are projected onto the neighboring cells are also considered a cluster. Subjects are defined as a cluster if the subjects are projected onto a cluster square area that consists of four adjacent cells in the SOM grid. Notice that subjects projected onto a cell in the SOM grid may be categorized into several different clusters. That is because the cell can belong to different cluster square areas in the SOM grid. In other words, a cell can be located at the upper left corner of a square area (Square I), the upper right corner of a square area (Square II), the lower left corner of a square area of (Square III), and the lower right corner of a square area (Square IV). Each of the four squares corresponds to a cluster. The semantic connections among the subjects at the second level may not be stronger than those at the first level. The U-matrix method adds a new dimension to subject clustering analysis. After an SOM display is generated, the map can be colored based on its corresponding U-matrix to facilitate clustering analysis. On the map an area with dark color represents low values in the U-matrix, while an area with light color represents high values in the U-matrix. A cell in the map with a high value implies that there are significant differences between this cell and its adjacent cells, and vice versa. When two groups of subjects are situated in two adjacent cells on the map and the colors of the two cells are different, it means that these two groups of the subjects are not relevant. If two groups of subjects are projected onto two adjacent cells and the colors of these two cells are the same, then the two groups of the subjects are relevant.

**Cluster Relevance Judgment**

The subjects identified in a cluster from the visual SOM analysis method are supposed to be semantically relevant, but this needs to be confirmed in the study. In an attempt to determine the perceived relevance among the health subject categories under investigation, we tested whether the perceived relevance among the health subject categories by the user’s traversal path analysis was really semantically related from the medical perspective. Toward this aim, we followed up on a user study based on the subject category data collected from the visual SOM analysis by asking a group of medical professionals to make a relevance judgment among the identified subjects in a cluster in different cases. The relevance judgment procedure is as follows.

The relevance between two subjects in a cluster was defined as a positive score between 0 and 1. That is, 0 stands for not relevant, 0.5 for relevant, and 1 stands for the most relevant. After two subjects were judged, a relevance score was given to them. For instance, if the relevance between two subjects was between relevant and most relevant, the relevance score 0.75 might be assigned. If they were most relevant, then relevance score 1 was assigned.

In a cluster, after the relevance judgments among all subjects were completed, a cluster relevance matrix (RM) was generated (see Equation (12)). In a cluster with \( n \) subjects, the relevance score between subject \( i \) and subject \( j \) is defined as \( c_{ij} \) (\( 0 \leq i, j \leq n \)). RM is a symmetric matrix because \( c_{ij} \) is always equal to \( c_{ji} \).

\[
RM = \begin{pmatrix}
c_{11} & \cdots & c_{1n} \\
\cdots & \cdots & \cdots \\
c_{n1} & \cdots & c_{nn}
\end{pmatrix}
\]

Since the relevance between a subject and itself is always equal to 1, we have the following equation:

\[
c_{ii} = 1, \quad 1 \leq i \leq n,
\]
Finally, the semantic accuracy of a cluster is defined in Equation (14):

$$SA = \frac{\sum_{i=1}^{n} \left( \sum_{j=1}^{n} \left( \frac{c_{ij}}{n} \right) \right)}{n}$$ (14)

In Equation (14), $\sum_{j=1}^{n}(c_{ij}/n)$, whose valid range is between 0 and 1, is the average relevance score between subject $i$ and all subjects in a cluster. $SA$ indicates the semantic strength among the subjects in a cluster. The denominator $(n)$ is used to normalize all average relevance scores in a cluster. Due to the normalization, the valid $SA$ value is also between 0 and 1. It is clear that the greater the $SA$ score of a cluster, the more semantically relevant the subjects in the cluster, and vice versa.

If there are $n$ subjects in a cluster, $n \times (n-1)/2$ relevance scores are generated and filled in the matrix described in Equation (12). From these raw relevance scores a $SA$ score of that cluster is finally calculated in Equation (14).

An analysis of variance (ANOVA) was used to examine whether there were significant differences among the three methods (the cluster identification method in the same cell based on path length 1, the cluster identification method in the adjacent cells based on path length 1, and the cluster identification method in the same cell based on path length 2) in terms of cluster semantic accuracy. The ANOVA analysis results demonstrated which method was the most effective and which method was the most ineffective.

**Evaluation of the U-matrix Method**

After an SOM display was produced and the U-matrix method was applied to the SOM display, the map was colored. The U-matrix method can help people to distinguish whether two groups of the objects projected onto two adjacent cells are relevant or not. When the objects were projected onto the same cell in the SOM feature map, these objects were related regardless of the color in the cell. That is because all these objects were highly relevant to the weight vector of the cell. However, when two groups of objects were projected onto two adjacent cells in the SOM display, these two groups of the objects might be either highly related or not. In other words, the subjects that were identified as a subject cluster in neighboring cells in the SOM display might be highly relevant or not.

In order to examine whether the proposed U-matrix method can effectively distinguish the two groups of the subjects that were not highly relevant and located in two adjacent cells in the SOM display, the following analysis was conducted.

We consulted the generated SOM display for traversal path length 1 (Figure 2), identified the adjacent cells whose colors were different, and identified the subjects projected onto these cells. These subjects would be treated as a cluster ($C_i$) in the cluster analysis for neighboring cells because of their adjacent relationship if the U-matrix method was not considered. The cluster $C_i$ ($i = 1, 2, \ldots, r$) was a cluster in the SOM display and the subjects in the cluster were mapped onto two adjacent cells and the colors of the two cells were different in the SOM display. The parameter $r$ was the number of the clusters that meet the conditions in the SOM display. The cluster $C_1$ was split into two subclusters: one cluster ($C_{11}$) included all the subjects in the cells sharing one color, and another cluster ($C_{12}$) contained all the subjects in the cells in the other color.

The medical professionals were asked to make a relevance judgment on the subjects in $C_1$, $C_{11}$, and $C_{12}$, respectively. The resultant data were used to fill the corresponding cluster relevance matrices for $C_1$, $C_{11}$, and $C_{12}$. Then the semantic accuracy scores for cluster $C_1$, $C_{11}$, and $C_{12}$ were calculated based on Equation (14). Finally, an ANOVA test was conducted to examine whether there were significant differences among the three groups ($C_1$, $C_{11}$, and $C_{12}$) of semantic accuracy scores. If both clusters $C_{11}$ and $C_{12}$ outperform cluster $C_1$ in terms of cluster semantic accuracy, it suggests that the U-matrix method can effectively distinguish the two groups of the subjects that are not highly relevant and are located in two adjacent cells in the SOM display. Otherwise, the proposed U-matrix method fails to do what it is supposed to do.

**Results Analysis and Discussion**

**Web Log Data Description**

The HealthLink Web log investigated covers the data from January 1, 2006 to December 31, 2006. The log file size is 1.41 GB and the format is Combined Log Format. Since this study concentrated only on the HealthLink subject directory, only the users’ traversal activities on the directory were extracted and collected from the transaction log. As a result, the users’ browsing activities on the directory topic, 47 subjects, and 2,099 associated articles were collected.

**Traversal Analysis for the Topic and Subjects in the Subject Directory**

Nodes in a subject directory can be classified into three basic categories: root, branch nodes, and leaves. In our case, the root was the topic, subjects were branch nodes, and articles were leaves. Notice that the number of columns in the proximity matrix was much smaller than that of the rows. That is because those columns whose cell values were equal to 0 were discarded to reduce computing complexity and enhance efficiency. Removal of these columns did not affect ultimate analysis results. Because the different variables (columns) had different value ranges, and the variables with large value ranges would dominate the generation of the SOM display, the input data had to be normalized with the ‘Var’ method (SOM_norm_variable, 2002). In the ‘Var’ method the normalization process is linear, and the variances of the variables are normalized to 1.

Sequential learning was selected because our pilot study showed it would better organize the input data and therefore
produce more balanced subject clusters. The newly defined U-matrix was employed to generate the background color for the SOM displays.

**Analysis for traversal path length 1.** To produce the visual configuration for the topic and subjects in the SOM space, a traversal matrix whose rows were only the topic and subjects was produced, and then the topic and subjects were projected onto the SOM space. Since the topic and subjects were included in this case, the matrix had 48 rows and 665 columns. The random initiation and sequential learning method were selected for the SOM display. Figure 2 shows the results. In Figure 2 the number labels (from 1 to 47) represent the subjects and the label 0 represents the topic. Table 1 shows the relationships between the label numbers and corresponding subjects in Figure 2.

The relevance among the subjects was reflected by their geometric vicinities in the visual configuration. That is, the subjects that were situated in the same grid were highly relevant, the subjects that were located in the adjacent grid cells were less relevant, and the subjects that were placed far away from each other were not relevant. In the resulting visual configuration, observe that some subjects were clustered in the same cells (such as {Pain (35), Physical medicine (36)} and {Brain nervous system (6), Mental health (29), Neurological disorders (31)}). Some subjects were mapped onto neighboring cells (such as {Drugs medications (13), Immune disorders (23)}), and other subjects were separated (such as {Digestive disease (12)} and {Hearing disorders (20)}). The clusters of the subjects in a single cell are summarized in Table 2. In Table 2, a number in parentheses behind a subject stands for the label of that subject in the visual configuration.

In the resulting visual configuration, 11 clusters projected onto the single cells were identified (Table 2), and 28 subjects were involved in the clusters. The average cluster size was 2.5 subjects. The largest cluster contained four subjects, and the smallest cluster had two subjects. Because subject clustering

<table>
<thead>
<tr>
<th>Cluster no.</th>
<th>Subject clusters identified in the same cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arthritis(4), back-problems(5), immunization(24), travel-medicine(43)</td>
</tr>
<tr>
<td>2</td>
<td>Pain(35), physical-medicine(36)</td>
</tr>
<tr>
<td>3</td>
<td>Brain-nervous-system(6), mental-health(29), neurological-disorders(31)</td>
</tr>
<tr>
<td>4</td>
<td>Eye-care(17), genetics(19), organ-transplants(34)</td>
</tr>
<tr>
<td>5</td>
<td>Cholesterol(9), heart-disease(21), high-blood-pressure(22)</td>
</tr>
<tr>
<td>6</td>
<td>Allergies(2), children’s health(8)</td>
</tr>
<tr>
<td>7</td>
<td>Emergency-medicine(14), infections(25)</td>
</tr>
<tr>
<td>8</td>
<td>Environmental-health(16), safety(40)</td>
</tr>
<tr>
<td>9</td>
<td>Preventive-medicine(37), wellness-lifestyle(46)</td>
</tr>
<tr>
<td>10</td>
<td>Diabetes(11), weight-control(45)</td>
</tr>
<tr>
<td>11</td>
<td>Kidney-disease(26), nutrition-and-herbs(32), women’s health(47)</td>
</tr>
</tbody>
</table>
That is because neighboring cells usually consist of an area instead of a simple cell in the SOM display. The subjects that appeared frequently in a cluster contained seven subjects, and the smallest cluster had two subjects.

Compared with Table 2, Table 3 has more subjects involved in a cluster, and therefore the average cluster size was larger. The largest cluster contained 15 subjects, and the smallest cluster had two subjects.

Compared with Table 2, Table 3 has more subjects involved in a cluster, and therefore the average cluster size was larger. That is because neighboring cells usually consist of an area instead of a simple cell in the SOM display.

Analysis for traversal path length 2. Because there were no direct links between subjects in the HealthLink subject directory, users could not jump from a subject node to another. However, browsing a directory usually requires a series of traversal activities to satisfy the user’s information need, for instance, drilling down the directory to narrow down a search scope, or selecting a related sibling subdirectory to explore other possibilities. These traversal activities are relevant to each other to some degree. If the user’s traversal path length increases to 2, more relevance information can be included and revealed, especially information on the relevance of one subject to another, which cannot be demonstrated in the visual configuration based on the traversal path length 1, due to the limitations of the directory structure.

The traversal matrix based on path length 2 had 2,006 rows and 1,725 columns after some zero columns were removed. The reason for the zero column removal is the same as that in the traversal matrix based on path length 1. The reserved topic and subjects were displayed in the SOM visual space. To make the generated figure comparable with Figure 2, the size of the SOM display space for traversal path length 2 was adjusted in the same way as the matrix based on path length 1. (See Figure 3 for the SOM display for traversal path length 2.) In Figure 3, since many subjects were projected onto the four cells in a red rectangle on the right side, which reduced the visibility of the projected subjects, the overloaded rectangle area was enlarged and displayed separately on the right side of the figure.

By using the same analysis method, the subjects clustered in a single cell are summarized in Table 4.

In the resulting visual configuration, seven clusters were identified (Table 4), and 34 unique subjects were involved in these clusters. The average cluster size was 4.9 subjects. The largest cluster contained 15 subjects, and the smallest cluster had two subjects.

Compared with the results for traversal path length 1 (Table 2), more subjects were involved in the clusters and the clusters became larger for traversal length 2 (Table 4). Since in Table 4 the clusters were formed in a single cell, no subject appeared in more than one cluster.

In order to compare the clustering methods, an experimental study was conducted. The significance level ($p$) or sig is 0.05 for all tests in this study. In other words, if $p$ or the sig is smaller than 0.05 in a test, the finding of the test is statistically significant.

Ten medical professionals participated in this experimental study. Among these 10 medical professionals, three were registered nurses and seven were MDs. All the participants were required to read instructions that include the purpose of the study, the definition of a task sheet, and the procedure of filling the sheet. Each of the clusters identified in Tables 2, 3, and 4 are relevant and can also be employed to form a cluster.

U-matrix analysis method was not applicable.

In addition, the subjects located in adjacent cells in the SOM display were also related to some degree. These subjects are relevant and can also be employed to form a cluster. Subject clusters in neighboring cells were identified from the SOM display and are summarized in Table 3.

Compared with the results for traversal path length 1 (Table 2), more subjects were involved in the clusters and the clusters became larger for traversal length 2 (Table 4). Since in Table 4 the clusters were formed in a single cell, no subject appeared in more than one cluster.

In order to compare the clustering methods, an experimental study was conducted. The significance level ($p$) or sig is 0.05 for all tests in this study. In other words, if $p$ or the sig is smaller than 0.05 in a test, the finding of the test is statistically significant.

Ten medical professionals participated in this experimental study. Among these 10 medical professionals, three were registered nurses and seven were MDs. All the participants were required to read instructions that include the purpose of the study, the definition of a task sheet, and the procedure of filling the sheet. Each of the clusters identified in Tables 2, 3, and 4 are relevant and can also be employed to form a cluster.

**TABLE 3. Subject clusters in neighboring cells.**

<table>
<thead>
<tr>
<th>Cluster no.</th>
<th>Subject name (label)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Arthritis(4), back-problems(5), immunization(24), travel-medicine(43), musculoskeletal(30)</td>
</tr>
<tr>
<td>2</td>
<td>Musculoskeletal(30), pain(35), physical-medicine(36)</td>
</tr>
<tr>
<td>3</td>
<td>Pain(35), physical-medicine(36), aging(1)</td>
</tr>
<tr>
<td>4</td>
<td>Aging(1), brain-nervous-system(6), mental-health(29), neurological-disorders(31), skin-disease(41)</td>
</tr>
<tr>
<td>5</td>
<td>Brain-nervous-system(6), mental-health(29), neurological-disorders(31), skin-disease(41), occupational-health(33)</td>
</tr>
<tr>
<td>6</td>
<td>Occupational-health(33), vitamins(44)</td>
</tr>
<tr>
<td>7</td>
<td>Eye-care(17), genetics(19), organ-transplants(34), cholesterol(9), heart-disease(21), high-blood-pressure(22)</td>
</tr>
<tr>
<td>8</td>
<td>Immune-disorders(23), skin-disease(41)</td>
</tr>
<tr>
<td>9</td>
<td>Skin-disease(41), alternative-medicine(3)</td>
</tr>
<tr>
<td>10</td>
<td>Allergies(2), children’s-health(8), emergency-medicine(14), infections(25), environmental-health(16), safety(40)</td>
</tr>
<tr>
<td>11</td>
<td>Environmental-health(16), safety(40), respiratory(39)</td>
</tr>
<tr>
<td>12</td>
<td>Respiratory(39), drugs-medications(13), immune-disorders(23)</td>
</tr>
<tr>
<td>13</td>
<td>Drugs-medications(13), preventive-medicine(37), wellness-lifestyle(46)</td>
</tr>
<tr>
<td>14</td>
<td>Preventive-medicine(37), wellness-lifestyle(46), feet(18)</td>
</tr>
<tr>
<td>15</td>
<td>Feet(18), alternative-medicine(3)</td>
</tr>
<tr>
<td>16</td>
<td>Feet(18), diabetes(11), weight-control(45)</td>
</tr>
<tr>
<td>17</td>
<td>Clinical-trials(10), liver(27), diabetes(11), weight-control(45)</td>
</tr>
<tr>
<td>18</td>
<td>Emergency-medicine(14), infections(25), sports-medicine(42), environmental-health(16), safety(40), public-health(38)</td>
</tr>
<tr>
<td>19</td>
<td>Diabetes(11), weight-control(45), kidney-disease(26), nutrition-and-herbs(32), women’s-health(47), cancer(7), liver(27)</td>
</tr>
<tr>
<td>20</td>
<td>Public-health(38), men’s-health(28)</td>
</tr>
<tr>
<td>21</td>
<td>Men’s-health(28), endocrine-system(15)</td>
</tr>
</tbody>
</table>

It is worth pointing out that the neighboring cell clustering analysis method can help identify some relevant subjects that cannot be identified in the single cell clustering analysis method, such as {Immune-disorders (23), skin-disease (41)}.
FIG. 3. The SOM display of the topic and subjects for traversal path length 2.

3, and 4 corresponds to a task sheet. A task sheet is an $n \times n$ matrix where $n$ is the number of the subjects in the cluster (see Equation (12)). The matrix was stored in an Excel file. Its initial status is $c_{ij} = 0$ and $c_{ii} = 1$ ($1 \leq i, j \leq n$). Both $c_{ij}$ and $c_{ii}$, which are cells in the matrix, represent a relevance score between two subjects in a cluster. The participants were required to complete each of the task sheets based on relevance among the subjects in the cluster. Since the matrix is symmetric, only half of the cells in the matrix need to be filled by the participants. After the raw data were collected from the participants, the unfilled empty cells in the cluster relevance matrix were filled and the semantic accuracy score for each of the clusters was calculated based on Equation (14). Then all the calculated semantic accuracy scores were placed into six categories for the two experimental tests. That is, the semantic accuracy scores from the same cell based on path length 1 (SAS1), the scores from the adjacent cell based on path length 1 (SAS2), the scores from the same cell based on path length 2 (SAS3), the scores from the light color cells split from the adjacent cells based on path length 1 (SAS4), the scores from the dark color cells split from the adjacent cells based on path length 1 (SAS5), and the scores from the adjacent cells based on path length 1 (SAS6). SAS1, SAS2, and SAS3 formed a group for one test while SAS4, SAS5, and SAS6 formed a group for another test. Finally,
the two groups were entered into SPSS (v. 16.00, 2007, Chicago, IL) for further data analysis.

Each participant examined 11 clusters identified in the same cells based on path length 1, 21 clusters identified in the adjacent cells based on path length 1, and 7 clusters identified in the same cells based on path length 2. Each participant made 24 subject relevance judgments for the clusters in the same cells based on path length 1, 129 subject relevance judgments for the clusters in the adjacent cells based on path length 1, and 128 subject relevance judgments for clusters in the same cells based on path length 2. As a result, 390 cluster relevance matrices were produced based on the relevance judgment results from the medical professionals. Each of the 390 matrices yielded a cluster semantic accuracy score. The 390 cluster semantic accuracy scores were categorized into three groups: Group I for the same cells based on path length 1, Group II for the adjacent cells based on path length 1, and Group III for the same cells based on path length 2. Finally, an ANOVA analysis method was applied to these three groups.

The descriptive results of the ANOVA analysis are displayed in Table 5. In the variable column, 1, 2, and 3 stand for Group I, Group II, and Group III, respectively. The greatest mean and smallest mean were 0.86116151 (Group I) and 0.75362664 (Group III), respectively. The greatest standard deviation and smallest standard deviation were 0.132696072 (Group I) and 0.161125469 (Group III), respectively. The greatest mean difference between Group I and Group II (0.107534865*) was also significant. In other words, the clusters from the same cells based on path length 1 outperformed the clusters from the adjacent cells based on path length 1, and the clusters from the same cells based on path length 1 also outperformed the clusters from the same cells based on path length 2. These were the factors that led to the significant differences among the three clustering analysis methods in the ANOVA analysis. Observe that there was a mean difference between Group II and Group III of 0.028436296. However, this mean difference was not significant. This means that the performance difference between the clusters from the adjacent cells based on path length 1 and the clusters from the same cells based on path length 2 was not significant.

The means for groups in homogeneous subsets are displayed in Table 8. In a Tukey HSD analysis, if the factors are grouped into a subset, these factors are not significantly different. In this test, two separate homogeneous subsets were produced (Table 8). Group II and Group III formed a homogeneous subset, and Group I formed a homogeneous subset. This confirms that there were significant differences between Group I and Group II (Group III) from a different perspective.

The visual display of the result can be seen in Figure 4. The Y-axis is the cluster semantic accuracy, and the X-axis is the visual clustering method. The factors 1, 2, and 3 stand for Groups I, II, and III respectively. As in all boxplots, the top of the box represents the 75th percentile, the bottom of the box represents the 25th percentile, and the line in the middle represents the 50th percentile.

The resultant data from the ANOVA show that the results of the clustering analysis in the same cell for path length 1 were better than both of these in the adjacent cells for path length 1, and those in the same cell based on path length 2 as well. This implies that as the path length increases, the semantic strength between the two subjects connected by the path

---

**TABLE 5. The descriptive data of the three clustering analysis methods.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std. deviation</th>
<th>Std. error</th>
<th>95% Confidence interval for mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lower bound</td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td>0.86116151</td>
<td>0.132696072</td>
<td>0.012652074</td>
<td>0.83608551</td>
</tr>
<tr>
<td>2</td>
<td>210</td>
<td>0.78206294</td>
<td>0.129260476</td>
<td>0.008919820</td>
<td>0.76447859</td>
</tr>
<tr>
<td>3</td>
<td>70</td>
<td>0.75362664</td>
<td>0.161125469</td>
<td>0.019258177</td>
<td>0.71520763</td>
</tr>
<tr>
<td>Total</td>
<td>390</td>
<td>0.79926884</td>
<td>0.141893515</td>
<td>0.007185057</td>
<td>0.78514244</td>
</tr>
</tbody>
</table>

---

**TABLE 6. ANOVA results of the three clustering analysis methods.**

<table>
<thead>
<tr>
<th></th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>0.629</td>
<td>2</td>
<td>0.315</td>
<td>16.908</td>
<td>0.000</td>
</tr>
<tr>
<td>Within groups</td>
<td>7.203</td>
<td>387</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7.832</td>
<td>389</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---
decreases, and the semantic strengths among the subjects in the same cells are stronger than those among subjects in the adjacent cells in the SOM display.

Now let us discuss the evaluation of the proposed U-matrix method. Since the U-matrix method was used to separate irrelevant clusters projected onto the adjacent cells in the SOM display, it was only applied to the cluster analysis for the adjacent cells based on path length 1 in this study. After consulting the generated SOM display (Figure 2) and Table 3, we identified four clusters in Table 3: C1 (No. 17) = (Clinical-trials(10), liver(27), diabetes(11), weight-control(45)), C2 (No. 18) = (Emergency-medicine(14), infections(25), sports-medicine(42), environmental-health(16), safety(40), public-health(38)), C3 (No. 19) = (Diabetes(11), weight-control(45), kidney-disease(26), nutrition-and-herbs(32), women’s health(47), cancer(7), liver(27)), and C4 (No. 20) = (Public-health(38), men’s health(28)). Within each of these clusters, the subjects were divided into two subclusters: the subcluster (SCL), whose subjects came from the light color cells, and the subcluster (SCD), whose subjects came from the dark color cells. In other words, although the subjects from C i (i = 1, 2, 3, and 4) were in a cluster, they came from the adjacent cells with different colors in the SOM display. For the cluster C1, SCL1 was {clinical-trials(10), liver(27)} and SCD1 was {diabetes(11), weight-control(45)}. For the cluster C2, SCL2 was {sports-medicine(42), public-health(38)} and SCD2 was {Emergency-medicine(14), infections(25), environmental-health(16), safety(40)}. For cluster C3, SCL3 was {cancer(7), liver(27)} and SCD3 was {Diabetes(11), weight-control(45), kidney-disease(26), nutrition-and-herbs(32), women’s health(47)}. For the cluster C4, SCL4 was {Men’s health(28)} and SCD4 was {Public-health(38)}. The cluster SCL i (i = 1, 2, 3, and 4) formed a relevance matrix and yielded a cluster semantic accuracy score based on Equation (14), as did the cluster SCD i (i = 1, 2, 3, and 4). As a result, there were 40 semantic accuracy scores in the Ci (i = 1, 2, 3, and 4) group, respectively. The greatest mean and smallest mean were 0.892500000 (the group SCLi) and 0.500000 (group SCLi and group Ci), respectively. The greatest mean and smallest mean in the minimum category were 0.1162154452 (the group Ci), respectively. The greatest standard deviation and smallest standard deviation were 0.1494435748 (the group SCDi) and 0.05845981 (the group Ci), respectively. The maximum values for all three groups were 1. The greatest value and smallest value in the minimum category were 0.5555500 (group Ci) and 0.500000 (group SCLi and group SCD), respectively.

The results of the ANOVA test for the U-matrix analysis are shown in Table 10. The sum of squares, degree of freedom, mean square, F-value, and p-value are listed in Table 10. The p-value (sig) was equal to 0.001, which is smaller than 0.05. This suggests that there were significant differences among group SCLi, group SCDi, and group Ci in terms of cluster semantic accuracy at the significance level (0.05).

Due to significant differences among the three groups in terms of cluster semantic accuracy, the Tukey HSD method was used to further determine the reasons for these differences. The Tukey HSD analysis results are presented in Tables 11 and 12. The factors (I and J) in the first column in the two tables were defined as follows: 1, 2, and 3 stand for group SCLi, group SCDi, and group Ci, respectively. Notice that the mean difference between group SCLi and group C1 (0.1098223000*) was positively significant, and the mean difference between group SCDi and group

### Table 7: Multiple comparisons of Tukey HSD for the three clustering analysis methods.

<table>
<thead>
<tr>
<th>(I) Factor</th>
<th>(J) Factor</th>
<th>Mean difference (I-J)</th>
<th>Std. error</th>
<th>Sig.</th>
<th>Lower bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.079098569*</td>
<td>0.016056842</td>
<td>0.000</td>
<td>0.04132058</td>
<td>0.11687656</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-0.079098569*</td>
<td>0.016056842</td>
<td>0.000</td>
<td>-0.11687656</td>
<td>-0.04132058</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>-0.107534865*</td>
<td>0.020858450</td>
<td>0.000</td>
<td>-0.15660992</td>
<td>-0.05845981</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>-0.028436296</td>
<td>0.018828316</td>
<td>0.000</td>
<td>-0.07273492</td>
<td>0.01586233</td>
</tr>
</tbody>
</table>

### Table 8: Homogeneous subsets for the three clustering analysis methods.

<table>
<thead>
<tr>
<th>Factor</th>
<th>N</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>70</td>
<td>0.75362664</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>210</td>
<td>0.78206294</td>
<td>0.86116151</td>
</tr>
<tr>
<td>1</td>
<td>110</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sig. 0.282 1.000
FIG. 4. The visual display of the ANOVA analysis for the three clustering analysis methods.

TABLE 9. The descriptive data of the U-matrix methods.

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std. deviation</th>
<th>Std. error</th>
<th>Lower bound</th>
<th>Upper bound</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>0.892500000</td>
<td>0.1217131980</td>
<td>0.0192445463</td>
<td>0.853574231</td>
<td>0.931425769</td>
<td>0.5000000</td>
<td>1.0000000</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>0.854450000</td>
<td>0.1494435748</td>
<td>0.0236291039</td>
<td>0.806655626</td>
<td>0.902244374</td>
<td>0.5000000</td>
<td>1.0000000</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>0.782677700</td>
<td>0.1162154452</td>
<td>0.0183752753</td>
<td>0.745510197</td>
<td>0.819845203</td>
<td>0.5555500</td>
<td>1.0000000</td>
</tr>
<tr>
<td>Total</td>
<td>120</td>
<td>0.843209233</td>
<td>0.1367168177</td>
<td>0.0124804808</td>
<td>0.818496635</td>
<td>0.867921832</td>
<td>0.5000000</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>

TABLE 10. ANOVA results of the U-matrix method.

<table>
<thead>
<tr>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>0.249</td>
<td>2</td>
<td>0.124</td>
<td>0.012</td>
</tr>
<tr>
<td>Within groups</td>
<td>1.975</td>
<td>117</td>
<td>0.017</td>
<td>0.012</td>
</tr>
<tr>
<td>Total</td>
<td>2.224</td>
<td>119</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>

C1 (0.0717723000*) was also positively significant. That is, both the clusters derived from the light color cells which were split from the adjacent cells based on path length 1 and the clusters from the dark color cells which were split from the adjacent cells based on path length 1 were more semantically accurate than the clusters from the adjacent cells based on path length 1. These two positively significant differences explain the results in the ANOVA. Observe that the mean difference between group SCLi and group SCDi was 0.0380500000. However, this difference was not statistically significant.

The means for groups in homogeneous subsets are displayed in Table 12. In this Tukey HSD test two separate homogeneous subsets were produced (Table 12). Group SCLi and group SCDi formed a homogeneous subset, as did group C1. This confirms that there were significant differences between group C1 and group SCLi (group SCDi) from a different perspective.

The visual display of the result can be seen in Figure 5. The Y-axis is the cluster semantic accuracy, and the X-axis is the cluster group identified from the U-matrix. The factors 1, 2, and 3 stand for group SCLi, group SCDi, and group C1, respectively. In Figure 5 the top of the box represents the 75th percentile, the bottom of the box represents the 25th percentile, and the line in the middle represents the 50th percentile.

The data from the ANOVA imply that the proposed U-matrix method can effectively distinguish the two irrelevant clusters which were adjacent to each other and situated in the cells with two different colors in the SOM display. Consequently, the clusters (No. 17, No. 18, No. 19, and No. 20 in
Table 3) should be split into two subclusters. Because cluster C₄ had only two subjects, each subcluster included only one subject after cluster splitting. Therefore, the two subclusters {public-health(38)} and {men’s health(28)} were no longer considered. Finally, cluster No. 20 was removed, and the three clusters No. 17, No. 18, and No. 19 in Table 3 were replaced by the six subclusters {clinical-trials(10), liver(27)}, {diabetes(11), weight-control(45)}, {sports-medicine(42), public-health(38)}, {Emergency-medicine(14), infections(25), environmental-health(16), safety(40)}, {cancer(7), liver(27)}, and {Diabetes(11), weight-control(45), kidney-disease(26), nutrition-and-herbs(32), women’s health(47)}. This shows that the U-matrix method enhanced the accuracy of the clustering analysis results for the neighboring cells in the SOM display.

The results in Tables 2–4 in conjunction with the results of the experimental study can be directly used to improve and adjust the subject directory. If a subject is in an identified cluster in the three tables, other subjects in the cluster can be added to the subject directory as terms related to the subject. For instance, cluster No. 5 in Table 2 consists of three related subjects: {cholesterol, heart-disease, high-blood-pressure}. For the subject cholesterol, the related terms heart-disease and high-blood-pressure can be added to the subject directory as the related terms of cholesterol. Similarly, cholesterol and heart-disease are added to the directory as the related terms of high-blood-pressure; and cholesterol and high-blood-pressure are added to the directory as the related terms of heart-disease.

Notice that the experimental study shows that the results of the clustering analysis in the same cell for path length 1 (Group I) were better than those in the adjacent cells for path length 1 (Group II), and these in the adjacent cells for path length 1 were better than those in the same cell based on path length 2 (Group III). As a result, the related terms from the different groups are ranked differently when they are added to the subject directory. In other words, the related terms from Group I are ranked higher than these from Group II, and the related terms from Group II are ranked higher than these from Group III. For instance, there are three clusters: {diabetes, weight-control}, {diabetes, weight-control, kidney-disease, nutrition-and-herbs, women’s health, cancer, liver}, and {weight-control, wellness-lifestyle}. They are selected from Group I, Group II, and Group III, respectively. It is evident that weight-control appears in all the three clusters. This implies other subjects in the three clusters can be added to the subject directory as the related terms of weight-control. And they are listed as terms related to weight-control in the subject directory as follows: diabetes, kidney-disease, nutrition-and-herbs, women’s health, cancer, liver, and wellness-lifestyle.

Because of the experimental results for the U-matrix method, the related terms in a cluster should be readjusted. The related terms in a cluster should form separate clusters if they are located in cells with different colors. For instance, public-health and men’s health cannot be treated as related terms in the subject directory. For the same reason, all the terms related to weight-control have to be revisited and revised. The two terms cancer and liver have to be removed from the related term list. Finally, the terms related to weight-control should be listed in the subject directory as follows: diabetes, kidney-disease, nutrition-and-herbs, women’s health, cancer, liver, and wellness-lifestyle.

Since different clustering methods can produce different cluster results, the evaluation method should not be based on a clustering method, but should be relatively independent of any clustering method to avoid possible bias. In the experimental study, medical professionals were used to evaluate cluster accuracy. These medical professionals had the medical expertise to make correct relevance judgments about the subjects in a cluster. And the evaluation method was not associated with any clustering method.

After an iterative learning process, the SOM algorithm reaches the converge condition. Consequently, the learning results which reflect the characteristics/features of the input dataset are preserved in the weight vectors associated with grid cells in the SOM display. The weight vectors include...
FIG. 5. The visual display of the ANOVA analysis for the U-matrix method.

rich information about the feature patterns in the dataset. It is the feature patterns that provide people with a subject semantic map about the dataset and present a semantic context in which the relationship among the projected objects (in this case, they are medical subject categories) in the SOM display can be visually analyzed. However, due to the high dimensionality of the weight vectors, the feature patterns cannot be observed and perceived by people in the SOM display. The purpose of the U-matrix method lies in the fact that it can effectively reduce the dimensionality of the weight vectors and reveal the feature patterns hidden in the weight vectors in a unique way. The U-matrix method can translate multidimensional information in the weight vectors into one-dimensional color information which is then represented in the cells of the SOM display. Different colors in the cells can form meaningful subject areas and help people distinguish irrelevant subjects that are located in adjacent cells.

Users’ traversal path lengths in a subject directory may vary in different browsing situations. The path lengths can range from 1, 2, 3, to a greater number. When the users’ traversal path data are used for subject cluster analysis, the association strength between the start subject and end subject play a crucial role in determining subject clusters. The association strength between two subjects decreases as the path length between the two subjects increases, as confirmed by the findings in this study. This suggests that the subject clusters identified in a shorter path length should be more accurate and sound than the subject clusters identified in a longer path length. In other words, although the traversal data based on path length 3 (or greater) can be employed to identify the subject clusters, the corresponding results may not be as sound as these based on path length 1 or 2.

Although there was no direct link between the subject nodes in the HealthLink directory, users did not have access related to subjects through the same directory root. In the leaves of the subject directory, the related subjects were listed for selection. From there, users could access related subjects. In other words, after users navigated the subject directory system and found a relevant article, they could select and access another related subject at will. In fact, if a user jumped from a subject node to another subject node through the root, it was excluded for the cluster analysis based on path length 1. That is because in that case the corresponding path length is 2 instead of 1.

Conclusion

Growing consumer health information on the Internet plays an extremely important role for the general public. In this study we employed the novel information visualization technique SOM in combination with a U-matrix method to analyze health subject clusters based on user traversal activities. The data investigated came from a transaction log of a consumer health-oriented portal (HealthLink). Traversal data on the health subject directory was extracted from the transaction log for analysis.

Subject clustering analyses for the same cells based on path length 1, the adjacent cells based on path length 1, and the adjacent cells based on path length 2 in the SOM display were
conducted. Subject clusters in the three cases were identified and analyzed.

An experimental study was also conducted to examine the effectiveness of the three clustering methods in the SOM display and the effectiveness of the proposed U-matrix method. In this study the relevance between two subjects in an identified cluster was determined by 10 medical professionals, and the semantic accuracy score for each of the clusters was calculated. The results from the experimental study demonstrated that the subjects in a cluster specified in the same cells based on path length 1 were more relevant (accurate) than those in a cluster specified in adjacent cells based on path length 2. The findings also showed that the subjects in a cluster identified from the same cells based on traversal path length 1 were more relevant (accurate) than the subjects in a cluster identified from the same cells based on traversal path length 2. There were no significant differences between the subjects in a cluster specified in the adjacent cells based on path length 1 and those in a cluster identified from the same cells based on traversal path length 2.

Notice that although there were no significant differences between the subjects in a cluster specified in the adjacent cells based on path length 1 and those in a cluster identified from the same cells based on traversal path length 2, the former still outperformed the latter to some degree due to the fact that the mean (0.78206294) of the clusters specified in the adjacent cells based on path length 1 is greater than that (0.75362664) of the clusters identified from the same cells based on traversal path length 2 (Table 5). In addition, the standard deviation (0.129260476) of the clusters specified in the adjacent cells based on path length 1 is smaller than that (0.161125469) of the clusters identified from the same cells based on traversal path length 2 (Table 5). This implies that the former method is more stable than the latter method in terms of the performance stability.

A new U-matrix was introduced and implemented in the SOM environment in this study. The new U-matrix considered the impact of both immediate neighboring cells and corner neighboring cells of a cell in the SOM display. The degree to which the corner neighboring cells have an impact on the cell is also taken into account in the newly defined U-matrix. The four clusters identified in neighboring cells based on path length 1 were specified and split by using the U-matrix analysis method. The findings in this study show that the U-matrix method enhanced the accuracy of the clustering analysis results for the neighboring cells in the SOM display. It is worth pointing out that when each of the qualified clusters was split into two subclusters based on their different colors in the SOM grid, both the subclusters significantly outperformed the combined clusters in terms of the cluster accuracy. This provides convincing evidence about the effectiveness of the proposed U-matrix method.

This research method can also be applied to Web subject directories of other domains, such as business, education, engineering, etc.

In future research, we hope to narrow down the associated articles within a specific subject and then visualize them in the SOM display. The visual configurations for traversal path length 3, 4, etc., can also be produced for clustering analysis as another future research topic.

Acknowledgment

We thank HealthLink for access to the transaction log data.

References


Ohura, Y., Takahashi, K., Pramudiono, I., & Kitsuregawa, M. (2002). Experiments on query expansion for Internet yellow page services using Web
log mining. In Proceedings of the 28th International Conference on Very Large Data Bases (pp. 1008–1018). Hong Kong.


