PROJECT MUSE®

Visual component plane analysis for the medical subjects based on a transaction log / L'analyse visuelle selon les plans de composants (*component plane analysis*) dans le cas de sujets médicaux à partir d'un journal de transactions

Jin Zhang, Lu An

Canadian Journal of Information and Library Science, Volume 34, Number 1, March/mars 2010, pp. 83-111 (Article)

Published by University of Toronto Press
DOI: *https://doi.org/10.1353/ils.0.0006*

For additional information about this article
https://muse.jhu.edu/article/382176

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

Visual component plane analysis for the medical subjects based on a transaction log

L'analyse visuelle selon les plans de composants (*component plane analysis*) dans le cas de sujets médicaux à partir d'un journal de transactions

Jin Zhang
International Collaborative Academy of Library and Information Science
Wuhan University, Wuhan, Hubei, 430072, P.R. China
School of information Studies, University of Wisconsin Milwaukee
Milwaukee, WI 53211
jzhang@uwm.edu

Lu An
School of Information Management, Wuhan University
Wuhan, Hubei, 430072, P.R. China
whuanlu@yahoo.com.cn

Résumé : Les trajets de navigation parmi des pages web par les utilisateurs peuvent jusqu'à un certain point révéler des associations entre ces pages. Deux méthodes de visualisation ont été utilisées : la méthode des « self-organizing maps (SOM) » et la méthode « component plane analysis » pour examiner les associations au sein de deux groupes de sujets médicaux reposant chacun sur des activités de navigation propres. Les sujets tirés d'un répertoire de sujets médicaux ont été extraits et les données de trajet au sein de ces sujets ont été analysées à partir d'un journal de transactions dans un portail consacré à la santé pour les consommateurs. Deux thèmes centraux ont été sélectionnés et deux groupes de sujets ayant un rapport avec ces thèmes ont été identifiés dans la visualisation SOM. Les jugements d'association sur les sujets ayant un rapport avec un sujet spécifié ont été faits dans le contexte des plans de composants de sujets. Les termes reliés aux deux thèmes centraux (nerfs et cœur) et le degré d'étroitesse de ce lien avec les deux thèmes centraux ont été identifiés dans l'étude. Cette étude montre que les activités de navigation des utilisateurs dans le répertoire de sujets sur la santé peuvent révéler et permettre d'identifier les associations entre les sujets dans le répertoire, et que l'analyse par plan de composants est fiable quand on l'utilise pour identifier les relations entre sujets dans ce genre de situation. Les résultats de cette étude peuvent être utilisés pour enrichir les thésaurus

© *The Canadian Journal of Information and Library Science*

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

médicaux, pour optimiser les répertoires de sujets médicaux, et pour comprendre le comportement informationnel des utilisateurs d'information médicale.

Mots-clés : self-organizing map, component plane analysis, journal de transactions web, répertoire de sujets médicaux, analyse de trajets informationnels

Abstract: Users' traversal activities among webpages can reveal associations among these webpages to some degree. Both the self-organizing map (SOM) method and component plane analysis method were used to investigate associations among the two groups of medical subjects respectively based on browsing activities. The subjects from a health subject directory were extracted and path data among these subjects were analyzed from a consumer health portal transaction log. Two focus themes were selected and two related subject clusters were identified respectively in the SOM display. The association judgements about the subjects related to a specified subject were made in the contexts of the subject component planes. Terms related to the two focus subjects (nerve and heart) and the degree to which these terms were related to the two focus subjects were identified in the study. This study shows that users' navigating activities on the health-related directory can reveal and identify the associations among the subjects on the directory, and the component analysis is reliable when it is applied to identify the subject relationships in such a situation. The findings of this study can be used to enrich medical thesauri, to optimize the medical subject directories, and to understand users' medical information-seeking behaviour.

Keywords: self-organizing map, component plane analysis, Web transaction log, health subject directory, traversal path analysis

Introduction

People retrieve medical information to look for an answer to a medical question, understand a medical concept or symptom, check properties for new drugs, etc. Medicine is a sophisticated and complicated area that involves a large medical vocabulary. It is widely recognized that there is a huge divide between medical information professionals and health consumers in use of medical information (Zhang, Wolfram, et al. 2008; Zhang and Wolfram 2009). A study on medical information-seeking behaviour from the health consumers' perspective would improve communication between medical professionals and common health consumers.

Consumer health portals are not only important repositories where people can tap into health information but also important channels where researchers can study common health consumers' information-search behaviour. A portal usually maintains a Web transaction log on

its server. A Web transaction log faithfully records all users' activities on the portal, ranging from users' traversal paths to submitted queries. A Web transaction log provides researchers with a unique way to study health consumers' information-seeking behaviour.

Users' browsing activities in a health subject directory may reveal the relevance among the subjects in the directory. *Relevance* in this study is defined as associations among the related subjects on the health subject directory. The association strength is determined primarily by the user's traversal activities among the subjects. The application of the self-organizing map (SOM) and component plane analysis methods to identifying the related subjects and association judgements among the involved subjects is a unique way to address this issue. An SOM is an artificial neural network that employs an unsupervised learning algorithm to convert input data in a high-dimensional space to a low-dimensional (two- or three-dimensional) presentation and preserves features of the data in the presentation. The component plane is a derived display of SOM, which reflects the contribution of the corresponding attribute to the overall SOM display.

The objectives of this study are to use the SOM approach and the component plane analysis approach to investigate the associations among health-related subjects in a consumer health portal transaction log, to identify the subject themes of interest, to analyze medical subject relationships from the users' perspective in the visual display, and to specify the related subjects and associations to a medical topic of interest in a component plane. The research question of this study is whether the associations among the related subjects revealed from the component plane analysis method based on user's path data on a health-related directory are accurate and sound. The findings of this study will make a positive contribution to understanding of health consumers' information-seeking behaviour, enriching medicine-related thesauri, optimizing the content organizations of the health information on a portal, and enhancing information retrieval.

Related research

Many studies have been done in relevance analysis. Users' query-term relevance was judged by the document descriptors to examine the consistency between query terms and descriptors (Fisk et al. 2003). In order

to test whether a novel method for organizing search results is more effective than the relevance-ranking method and clustering method, 15 patients with breast cancer and their family members were invited to retrieve relevant medical information using three methods: a ranking tool, a cluster tool, and a category tool. The results showed that the category tool induced significantly more answers in a certain period and significantly more satisfactory search experience than the other two tools (Pratt and Fagan 2000). In another study, content-similarity networks were constructed to improve effectiveness of an information retrieval system (Lin 2008).

Medical-term analysis research has attracted attention of many researchers, such as their studies on similar terms (Tsuruoka et al. 2007), synonyms and hyponyms (Liu and Friedman 2003), and normalized terms (Tsuruoka et al. 2008). Wang et al. (2008) proposed a novel-feature-value schema, which ranges from a lower-level domain-independent "string feature" to a higher-level domain-dependent "semantic template feature." Pesquita et al. (2008) conducted a systematic evaluation of the gene ontology–based semantic similarity measures used for medical-term relevance analysis. Jelier et al. (2008) presented an online tool Anni, providing an ontology-based interface for MEDLINE users. The tool can be used to retrieve documents and present the associations of related biomedical concepts like genes, drugs, and diseases. In order to investigate the difference between the machine-generated terms and human-generated terms, manually created ontology terms were compared with automatically derived terms in four different automatic term-extraction methods (Alexopoulou et al. 2008). Mohanty et al. (2008) developed Common Data Elements for annotating mesothelioma specimens using controlled vocabulary, ontology, and semantic modelling methodology to improve both syntactic and semantic interoperability.

In recent years, research on Web transaction log analysis has increased dramatically. In general, transaction log analysis of users' information-seeking behaviour can be divided into two aspects: users' query analysis and users' traversal path analysis. Users' query analysis usually focuses on submitted query terms and their concurrences, whereas traversal path attempts to discover users' access patterns on the basis of navigating activities. There are many traversal-path analysis algorithms, such as Apriori-like sequential-pattern mining techniques (Adami, Avesani, and Sona 2003), Web access pattern tree (WAP-tree) (Pei et al. 2000), and

conditional sequence base mining algorithm (Zhou, Hui, and Fong 2006), etc.

In one recent study (Zhang et al. 2009), users' traversal activities on a health-subject directory were investigated, and the relationships among subjects on the directory were visualized and analyzed in the SOM environment. The findings show that the semantic relationships among the subjects can be identified on the basis of users' browsing activities in the health-subject directory by using the SOM method. In this study, the associations among the related subjects based on users' traversal activities in a health-related directory are investigated. It is clear that the objects of the previous studies on medical relevance analysis have focused on users' query terms, medical thesaurus terms, medical descriptors, and medical documents. The relevance among those objects has been measured by query-term frequencies, semantic similarities, or human judgements. A subject directory is used as an information navigation system. Associations among the subjects in the directory provide rich information. The subjects on the health-related subject directory, which are seldom studied as the objects of relevance analysis, are different from users' search terms, medical thesaurus terms, medical descriptors, and medical documents. Therefore, the relevance between the subjects on the health-related subject directory should be measured differently.

Research methods

The HealthLink subject directory

In this study, the investigated transaction log came from HealthLink (http://healthlink.mcw.edu/).[1] The HealthLink portal incorporates a medical subject directory that guides health consumers in browsing health information in the portal. The subject directory is a three-level hierarchical structure, whose root is called the topic. The 47 nodes associated with the directory branches are medical subjects such as allergies/asthma, alternative medicine, and so on. Table 1 shows a complete list of all subjects in the directory. Labels in the table stand for the corresponding subjects and are displayed in the later visual displays. The topic is not listed in table 1. Its label number is 0. All health-related articles classified under the corresponding subjects are located at the lowest level of the subject directory. Users can browse the collection by starting with the directory root "Browse by Topic" (http://healthlink.mcw.edu/topics/),

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

Table 1. HealthLink subject directory

Label	Subject	Label	Subject
1	Aging	25	Infections
2	Allergies	26	Kidney disease
3	Alternative medicine	27	Liver
4	Arthritis	28	Men's health
5	Back problems	29	Mental health
6	Brain nervous system	30	Musculoskeletal
7	Cancer	31	Neurological disorders
8	Children's health	32	Nutrition and herbs
9	Cholesterol	33	Occupational health
10	Clinical trials	34	Organ transplants
11	Diabetes	35	Pain
12	Digestive disease	36	Physical medicine
13	Drugs medications	37	Preventive medicine
14	Emergency medicine	38	Public health
15	Endocrine system	39	Respiratory
16	Environmental health	40	Safety
17	Eye care	41	Skin diseases
18	Feet	42	Sports medicine
19	Genetics	43	Travel medicine
20	Hearing disorders	44	Vitamins
21	Heart disease	45	Weight control
22	High blood pressure	46	Wellness lifestyle
23	Immune disorders	47	Women's health
24	Immunization		

then jumping to a subject of interest, and finally reaching related full-text articles (see figure 1).[2]

Data processing and cleaning

The investigated HealthLink Web log covers data collected between 1 January and 21 December 2006. The log file size is 1.41GB and the format is Combined Log Format. A total of 465,289 records were parsed and useful fields were extracted, such as IP address of host, time and

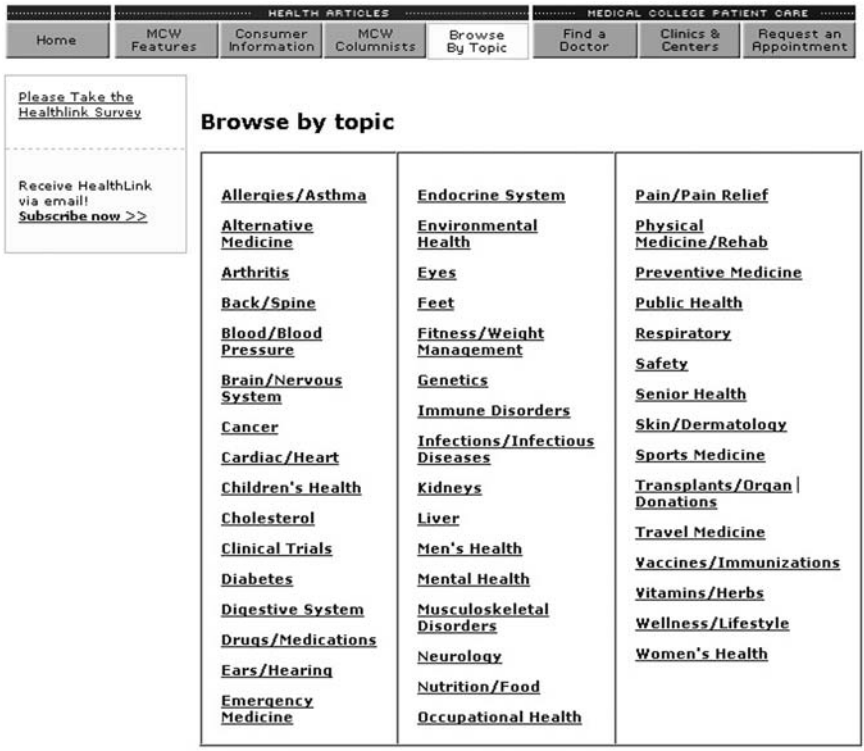

Figure 1: Display of the HealthLink directory

time zone, method, destination page, HTTP version, response code, bandwidth, referrer, and user agent. In this study, all users' traversal activities from a subject (or the topic) to another subject (or a topic or an article) were preserved and extracted for analysis. In order to identify a traversal path from the Web transaction log, all referrers (which are the 47 subjects or the topic) and destinations (which are the topic, or 47 subjects, or articles) had to be specified, extracted, and recorded for later data processing and analysis. Other activities that did not contribute to these traversal paths, such as queries or downloads, were excluded. A Web log analysis software package Web Log Explorer was used for these analyses.

In order to protect the privacy of the HealthLink transaction users, all personal data such as users' IP addresses and user IDs were stripped out prior to data analysis. In other words, users' privacy was well protected. In fact, unlike a transaction log analysis for users' query terms, the tran-

saction log analysis for users' browsing activities does not handle keywords from users. It significantly reduces the risk of revealing users' identifiers and other sensitive issues in the study.

Construction of the SOM input matrix

The users' traversal activities were presented in an $m \times n$ matrix. This matrix was used as the input data format for the later SOM and component plane analysis. The rows (m) of the matrix stand for the objects that are analyzed and visualized in the SOM display space, while the columns (n) of the matrix define the attributes of the objects.

The method of constructing the SOM input matrix (MP) is described as follows. Suppose a matrix (see equation 1) has m rows and n columns. Rank all the subjects and the topic alphabetically and number them from 1 to m as the rows. Rank all the investigated webpages (including the topic, subjects, and articles), and number them from 1 to n as the columns. Here cij ($I = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n$) stands for a cell of the matrix. If there are traversal activities from webpage i to webpage j, the element cij defines the number of the traversal activities from webpage i to webpage j. For instance, if the path data from webpage i to webpage j occur k times, then cij equals k. If there is no travel from webpage i to webpage j, then cij equals zero. It means that only the webpages that have at least one navigating activity are included in the traversal data analysis.

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

In fact, equation 1 defines a traversal activity matrix for path length one.[3] In the same way a traversal activity matrix for path length two can be defined as follows: if the traversal activities from webpage i to webpage j via another webpage r occur k times, then cij equals k. Here the webpage r is an intermediate webpage. It can be any webpage.

Selection of the two focus themes

In order to determine the study themes, the SOM display based on traversal path length one (see figure 2) was produced. In the SOM process, the linear initiation and sequential learning methods were selected. In the

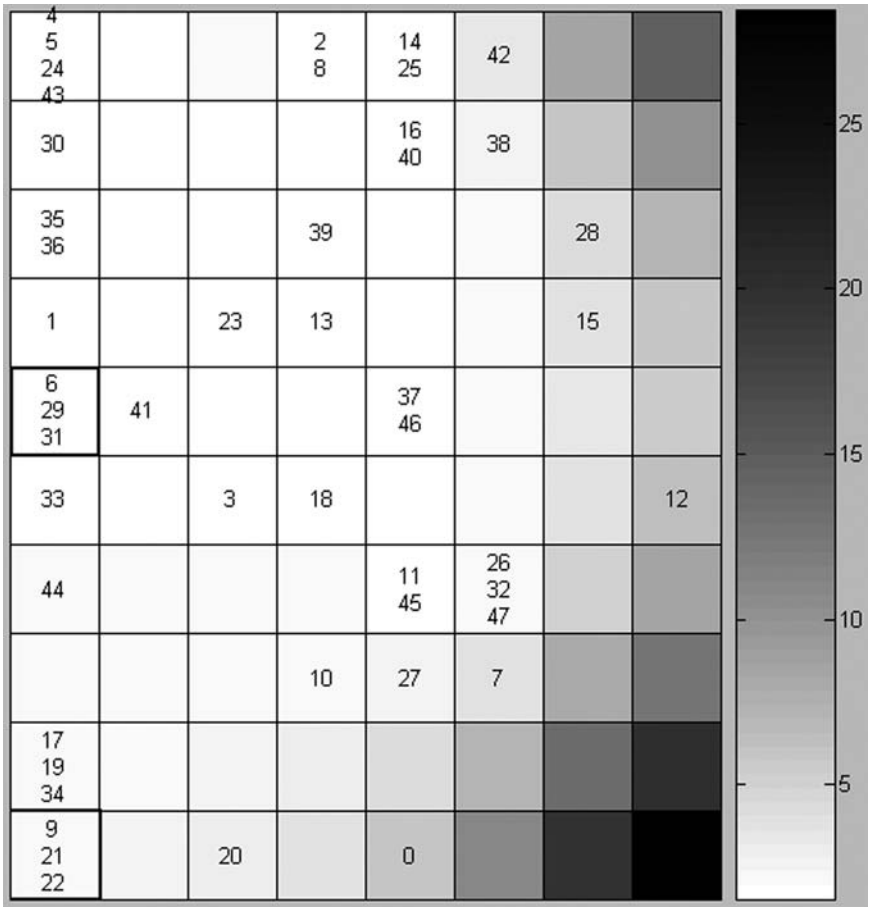

Figure 2: SOM display for traversal path length one
Note: Dark cells represent high values and the light cells represent low values in weight vectors. Numbers in the grid represent the projected subjects.

resultant SOM display, the two themes marked by the highlighted squares on the left were identified: one is heart and the other is nerve. Each theme corresponds to a group of related subjects. Because the subjects in each group were projected onto the same cell in the SOM display, they were highly related to each other. Therefore each theme contains three highly related subjects.

1. The nerve theme group: brain nervous system (6), mental health (29), and neurological disorders (31).

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

2. The heart theme group: cholesterol (9), heart disease (21), and high blood pressure (22).

The number within parentheses represents a label in the SOM display in figure 2. See table 1 for the relationship between a subject and a label number.

Analysis method

Component planes are significant visual representations in the SOM literature (Kohonen 1995). Each component plane is generated by assigning a colour level proportional to a specified weight vector value to each node in the SOM grid. A component plane reveals and demonstrates the distribution of the values of the specified component (the attribute) in a categorized fashion and has been successfully used by many researchers as a standard visualization data analysis technique in a wide variety of fields (Abonyi et al. 2003; Díaz et al. 2003; Laine 1998; Postolache et al. 2005). Ersoya et al. (2007) used the SOM technique to cluster the volcanic ash data collected from different fragmentation mechanisms and conducted the component plane analysis for each component to observe its discriminative performance. It was found in the study that the SOM technique was more accurate than conventional statistical methods in data-clustering analysis, and the component plane method was more reliable in determining the distinctive parameters. Fish and lamprey relative biomass data were categorized by using the SOM technique (Kruka et al. 2007), and the study showed that the SOM method provided more detailed information on the mutual relations between species through component planes than the detrended correspondence analysis (DCA) through points in a multivariate space.

Because of the complex nature of medicine, relationships among medical subjects or concepts are complicated. As a result, health consumers' activities on the portal are also complicated. As an effective approach for illustrating and analyzing complicated relationships and connections among objects, self-organizing maps (SOM) (Kohonen 1982) and the component plane analysis (Kohonen 1995) can reveal the relationships among medical subjects. Health consumers usually retrieve health-related issues such as diagnosis, injury, disease, therapy, symptom, disease prevention, nutrition. Each of these terms is intrinsically related to the others. It is

the relatedness of the health issues that makes health-information seeking complicated (health information seeking definitely includes the traversal activities on a health-related directory). A visualization data-analysis method like the component plane analysis method provides people with a two-dimensional space where objects are projected, multiple features of an object can be effectively preserved, and multiple connections among the objects can be displayed and analyzed in the visual space. Therefore, it is natural to conduct a health-subject analysis by using the component plane analysis method.

A subject directory is a hierarchical structure that shows semantic relationships among subject nodes. Since webpages are mapped onto the subject directory, the semantic relationships among webpages are illustrated in the directory. When users browse a subject directory, they usually select a related subject or node and jump to a related node. This implies that users tend to traverse among subjects that are related to each other. Therefore users' browsing paths in a subject directory can reveal semantic relationships among subjects in a directory. In other words, if users' traversal information on a subject directory is collected and the traversal data in the directory are analyzed by using a clustering analysis method like the SOM technique, then user-oriented semantic relationships among the subjects can be demonstrated and visually analyzed in the SOM display.

Furthermore, a component plane analysis for a specified subject can be conducted for more detailed analysis (Kohonen 1995). Every cell in the SOM display has an associated weight vector. A weight vector keeps and records the learning results during the SOM display generation. Simply put, the weight vectors preserve data features of a dataset. A weight vector has the same data structure as an input raw data vector. It plays a critical role in determining the final object configuration in the SOM display. In the component plane analysis for a specified subject, the corresponding attributes in all weight vectors are extracted and separated from the original SOM display to yield a new display whose size is exactly the same as that of the SOM display. That is, both the component plane and the SOM display share the same display structure. But the difference between this newly yielded component plane and the SOM display is that a cell in the component plane corresponds to a single attribute value while a cell in the SOM display corresponds to a weight vector. If the weight vector has n elements (attributes), it means

that n component planes can be derived from the corresponding SOM display. For instance, the i^{th} attribute in all the weight vectors can be extracted from the SOM display to form a component plane for the i^{th} attribute ($1 \leq i \leq n$).

Since each cell in the component plane is associated with a single attribute value, the cell values can be easily converted into colours in such a way that a higher cell value corresponds to a darker colour and a lower cell value corresponds to a lighter colour. After colour conversion is complete, the component plane can be partitioned by different colours. Because the component plane has the same structure as the SOM space, the coloured component plane can be merged with the SOM display where all the subjects are projected. Consequently, the resulting subject clusters in the SOM display can be analyzed in the context of the coloured component plane.

In the component plane the higher a cell value (or the darker the colour), the greater the traffic from all subjects to the specified subject. In other words, the subject cluster located in a cell with a high value (or dark colour) tends to have more traffic to the specified subject than a subject cluster in a cell with a low value (or light colour). This means that subjects in a cell with a higher value in the component plane are more relevant to the specified subject than subjects in a cell with a lower value. Following this principle, subject clusters in the component plane can be ranked by associations to the specified subject on the basis of their colours. Four basic association categories are defined for association judgement: strongest association, strong association, some association, and weak association.

1. In order to make relevance judgements consistent in the component plane analysis for all subjects, the colour bar was divided into four parts. Red indicates that subjects projected onto these areas had the strongest association to the specified subject.
2. Orange indicates that subjects projected onto these areas had a strong association to the specified subject.
3. Yellow or green indicates that subjects projected onto these areas had some association to the specified subject.
4. Cyan or light blue indicates that subjects projected onto these areas had a weak association to the specified subject.

It is worth pointing out that the cell values of a component plane for the i^{th} subject attribute are affected by traversal frequencies from all subject nodes in the subject directory to the i^{th} subject node, but they are not dominated by the traversal frequencies only. Frequent browsing from one subject (S) to the specified i^{th} subject may lead to a high cell value of the subject (S) in the component plane, while infrequent navigating from the subject (S) to the specified i^{th} subject may result in a low cell value of the subject (S) in the component plane. However, the final cell value of the subject (S) in the component plane is also affected by other factors such as browsing from other subjects to the specified i^{th} subject.

In this study all subject component analyses were conducted on the basis of the SOM display for traversal path length two, because in the HealthLink subject directory there is no direct link between two subjects; the connection between two subjects is via a third webpage such as an article, and the traversal relationship between any two subjects had to be established on the basis of a browsing matrix for path length two.

In order to evaluate the results of the component plane analysis for association accuracy, the results were reviewed by a physician who was an experienced medical researcher with expertise in the field. After related subjects to a subject were identified in a component plane analysis, the association for each related term was judged by the physician against the same scale as described above. It is widely recognized that relevance is situational and related to the individual's information need. But in this case, the traversal analysis in the transaction log is not based on one individual's information need and traversal behaviour. Instead, the traversal analysis reflects a large group of users' information needs and accumulative navigation. Although information needs and navigational activities on the subject directory differ among individuals, as a whole a large group of users' information-seeking activities on the directory may reveal some relevance patterns among the subjects. As a result, each related subject was assigned an association score based on its association degree. Association scores one, two, three, and four were defined as the strongest association, strong association, some association, and weak association, respectively. It is clear that the association scale system is similar to that in the component analysis. Finally, a *t*-test was used to examine whether there was a significant difference between the results from the component plane analysis method and the results from the medical professional judgements. No significant difference implies that results from

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

the component plane are sound and reasonable. Otherwise, results from the component plane are questionable. Since there were two themes investigated in this study, two *t*-tests were conducted for each of the two themes respectively.

Findings

The research question of this study is whether associations among related subjects revealed from the visual component analysis method based on user's browsing on a health-related directory are accurate and sound.

Data description

Since this study concentrated on the HealthLink subject directory, only users' traversal activities on the directory were extracted and collected from its transaction log. Users' browsing data on the topic (the root of the directory) resulted in 47 subjects and 2099 associated articles.

Component plane analysis for selected subjects

Component plane analysis for nerve theme group

Component plane analysis for brain, nervous system

The component plane for the subject "brain, nervous system" was generated on the basis of the traversal matrix of path length two (see figure 3).

The legend on the right of the SOM display indicates that dark cells represent high values of the variable (subject) and light cells represent low values of the variable (subject) in the weight vectors. The values in the legend range from less than zero to more than five.

The subjects located in the relatively high-value cells were identified and ranked in a descending order against their colours or association strength to the subject "brain, nervous system" in table 2.

Component plane analysis for mental health

The resultant component plane for mental health is shown in figure 4. The value in the legend ranges from less than -0.5 to 2, which is narrower than that in figure 3.

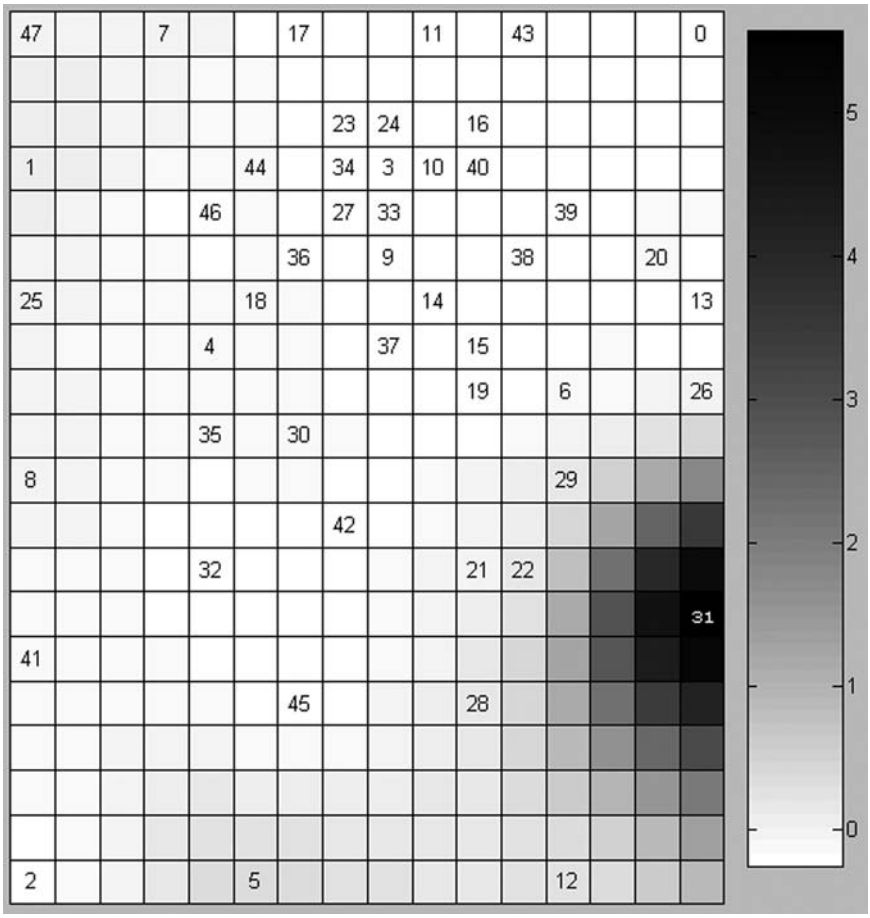

Figure 3: Component plane for brain, nervous system
Note: Dark cells represent high values and light cells represent low values in weight vectors. Numbers in the grid stand for the projected subjects.

Judged by the same criteria, subjects related to mental health were ranked in a descending order in table 2.

Component plane analysis for neurological disorders

The component plane for neurological disorders is shown in figure 5. The association strength order of the subjects identified by the component plane analysis was exhibited in table 2.

Table 2. Summary of the related subjects for the nerve theme group

Subject	Strongest association	Strong association	Some association	Weak association
Brain, nervous system	Neurological disorders (31)			Mental health (29), high blood pressure (22), digestive disease (12)
Mental-health	Children's health (8), brain, nervous system (6), neurological disorders (31)	Aging (1), women's health (47)	Men's health (28), mental health (29), drugs, medications (13), kidney disease (26)	Wellness lifestyle (46), nutrition and herbs (32), weight control (45), digestive disease (12), pain (35), genetics (19), hearing disorders (20)
Neurological disorders	Neurological disorders (31)			Brain, nervous system (6), infections (25), mental health (29), pain (35), back problems (5)

Note: Subject numbers in the figure are enclosed in parentheses.

Evaluation of results of nerve theme group

The significance level (p) at 95% confidence interval selected was .05 for the two t-tests in this study. In other words, if p is less than .05 in a test, the finding of the test is statistically significant. Otherwise, there is no significant difference.

The descriptive data of the first t-test are shown in table 3. Since the nerve theme group includes all subjects in table 2, there are 26 subjects.

Because the p-value in the test is greater than .05, there is no significant difference between the results from the component plane analysis method for the nerve theme and the results from the medical professional judgements. This implies that the results from the component plane are sound and reasonable.

Component plane analysis for heart theme group

Component plane analysis for cholesterol

The resultant component plane of cholesterol is shown in figure 6. According to the cell colours in figure 6, the subjects related to cholesterol are listed in association strength descending order in table 5.

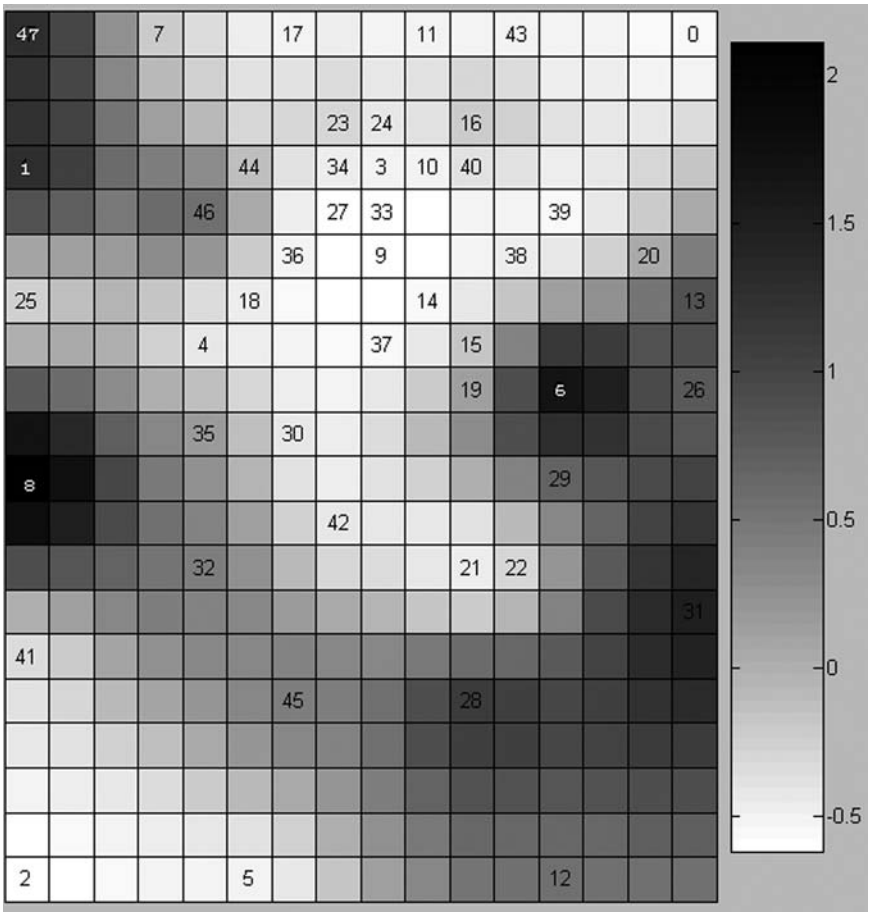

Figure 4: Component plane for mental health
Note: Dark cells represent high values and light cells represent low values in the weight vectors. Numbers in the grid stand for the projected subjects.

Component plane analysis for heart disease

The resultant component plane for heart disease is shown in figure 7. According to the cell colours, the subjects related to heart disease are summarized in table 5.

Component plane analysis for high blood pressure

The produced component plane for high blood pressure is shown in figure 8. According to the cell colours, the related subjects to high blood pressure are listed in association strength descending order in table 5.

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

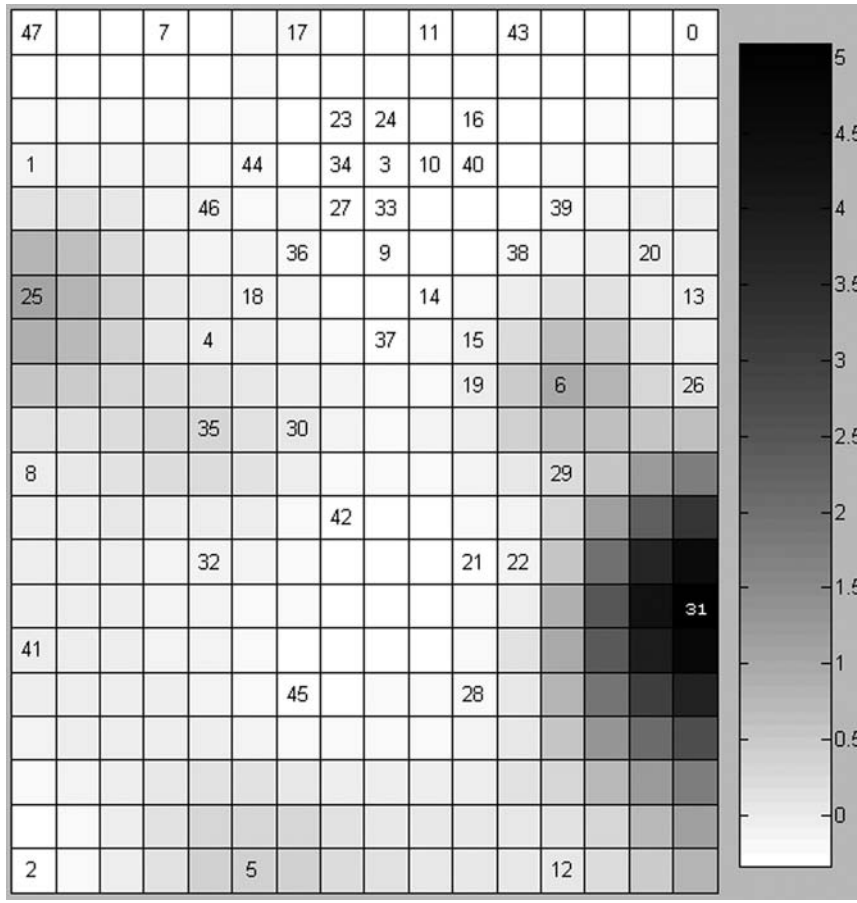

Figure 5: Component plane for neurological disorders
Note: Dark cells represent high values and light cells represent low values in the weight vectors. Numbers in the grid stand for the projected subjects.

Table 3. Descriptive data of the nerve theme group

Factor	*N*	Mean	Standard deviation	Standard error mean
Component plan analysis	26	2.5769	1.23849	.24289
Professional judgement	26	3.1154	1.21085	.23747

Table 4. *t*-test results of the nerve theme group

t	*df*	Sig. (2-tailed)	Mean difference	Standard error difference	95% confidence interval of the difference	
					Lower	Upper
−1.585	50	.119	−.53846	.33968	−1.22074	.14381

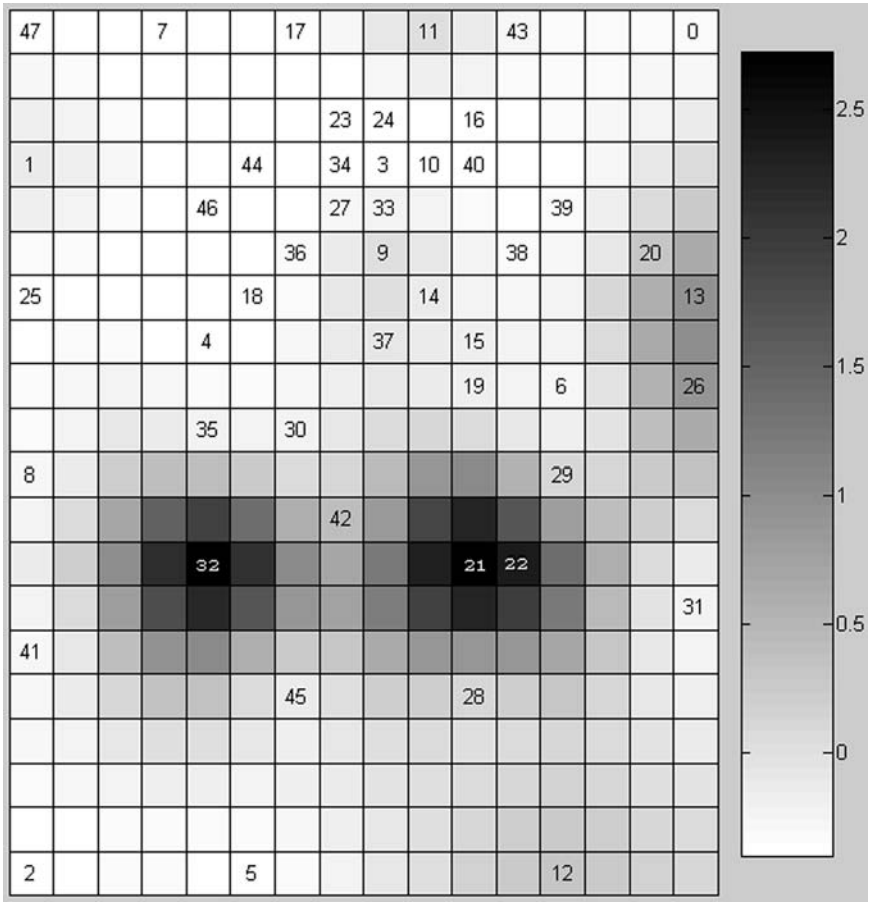

Figure 6: Component plane for cholesterol
Note: Dark cells represent high values and the light cells represent low values in the weight vectors. Numbers in the grid stand for the projected subjects.

Table 5. Summary of related subjects for heart theme group

Subject	Strongest association	Strong association	Some association	Weak association
Cholesterol	Nutrition and herbs (32), heart disease (21), high blood pressure (22)			Drugs, medications (13), kidney disease (26), sports medicine (42), digestive disease (12), hearing disorder (20)
Heart disease	Heart disease (21), high blood pressure (22)	Kidney disease (26)	Cholesterol (9)	Diabetes (11), emergency medicine (14), drugs, medication (13), nutrition and herbs (32), occupational health (33), sports medicine (42), digestive disease (12)
High blood pressure	Heart disease (21), high blood pressure (22)			Kidney disease (26), weight control (45)

Note: Subject numbers in the figure are enclosed in parentheses.

Evaluation of results of the heart theme group

The significance level (p) at 95% confidence interval of the difference is .05, which is the same as the previous test.

The descriptive data from the second t-test are demonstrated in table 6.

Final results of the t-test are shown in table 7. Because the p-value in the test is .223, which is greater than .05, there is no significant difference between results from the component plane analysis method for the heart theme and results from medical professional judgements. This suggests that the results from the component plane in this case are also sound and reliable.

Discussion

Although figures 3, 4, and 5 are the component planes for the nerve-related subjects, the colour distributions in the three figures are quite dif-

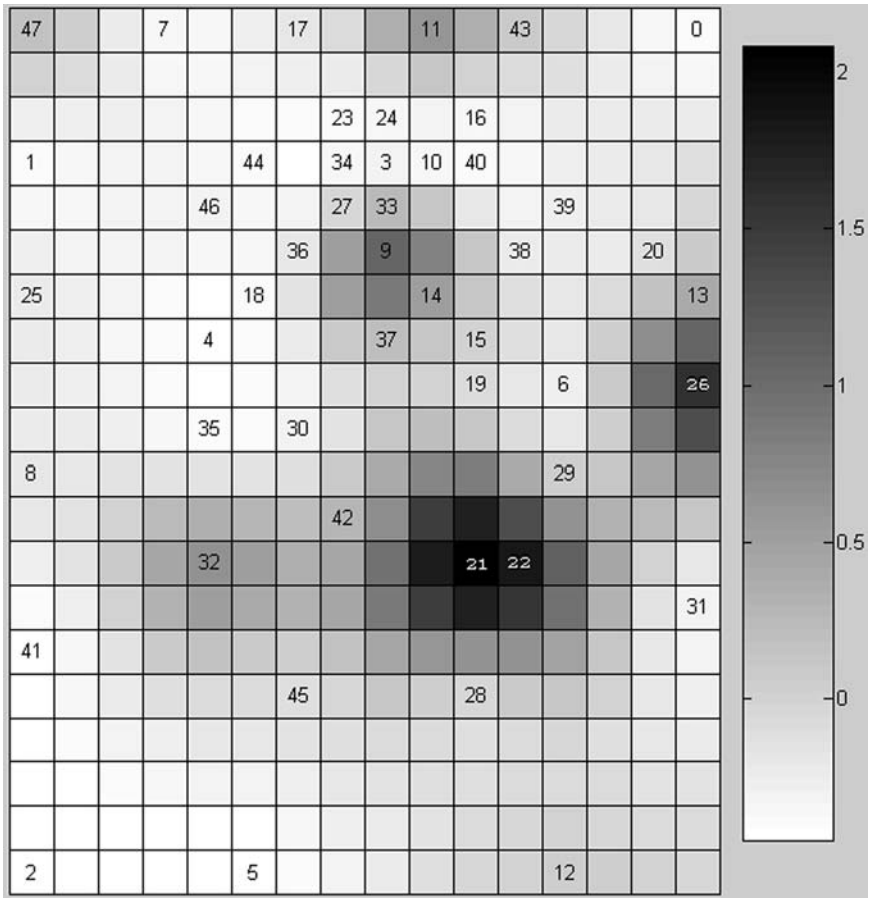

Figure 7: Component plane for heart disease
Note: Dark cells represent high values and the light cells represent low values in the weight vectors. Numbers in the grid stand for the projected subjects.

ferent. There is only one dark area in figure 3. No subject is situated in medium grey (orange or yellow) cells. This means that there is no strong associated or some associated subject in this component plane. However, it is interesting that there are three separate areas with darker areas in figure 4, although they are clearly not connected. Notice that mental health (29) is located in a cell of medium grey (green) rather than dark (red or yellow) in figure 4, although figure 4 is the component plane for mental health itself. As for figure 5, there is only one area with dark colour in it.

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

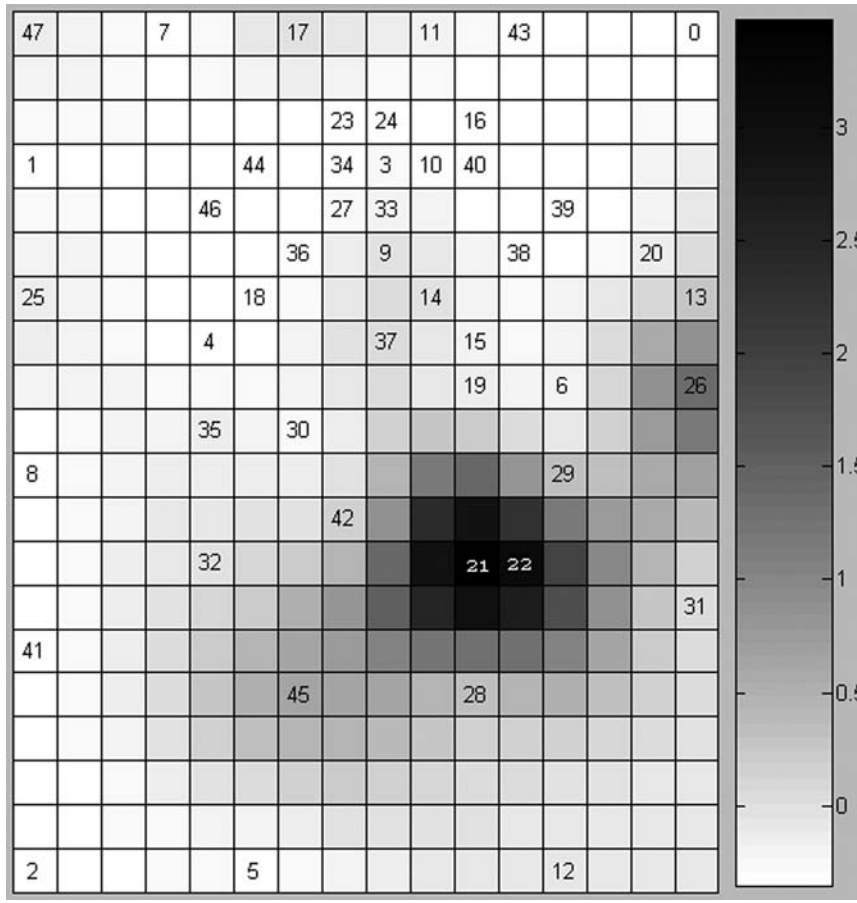

Figure 8: Component plane for high blood pressure
Note: Dark cells represent high values and light cells represent low values in the weight vectors. Numbers in the grid stand for the projected subjects.

Table 6. Descriptive data of the heart theme group

Factor	*N*	Mean	Standard deviation	Standard error mean
Component plane analysis	23	2.5217	.94722	.19751
Professional judgement	23	2.9565	1.39734	.29137

Table 7. *t*-test results of the heart theme group

t	*df*	Sig. (2-tailed)	Mean difference	Standard error difference	95% confidence interval of the difference	
					Lower	Upper
−1.235	44	.223	−.43478	.35200	−1.14419	.27463

Table 8. Summary of nerve-related subjects

	Strongest association	Strong association	Some association	Weak association
Neurological disorders	3			
Brain, nervous system	1			1
Children's health	1			
Aging		1		
Women's health		1		
Mental health			1	2
Men's health			1	
Drugs, medications			1	
Kidney disease			1	
Digestive disease				2
Infections				1
High blood pressure				1
Pain				1
Back problems				1
Wellness lifestyle				1
Nutrition and herbs				1
Weight control				1
Genetics				1
Hearing disorders				1

The nerve theme-related subjects and their association strength distribution are summarized in table 8, where the number in a cell indicates the frequency that the related subject is rated in the three subject component analyses for brain, nervous system, mental health, and neurological disorders. The row of the table lists all involved subjects in association judgement.

Table 9. Summary of heart-related subjects

	Strongest association	Strong association	Some association	Weak association
Heart disease	3			
High blood pressure	3			
Nutrition and herbs	1			1
Kidney disease		1		2
Cholesterol			1	
Drugs, medications				2
Sports medicine				2
Digestive disease				2
Diabetes				1
Emergency medicine				1
Weight control				1
Hearing disorders				1
Occupational health				1

Colour distributions in component planes for the heart theme are also different. There are three separate areas with light colours in figure 6, five in figure 7, and two in figure 8.

Table 9 summarizes all involved subjects and judgements in the three component plane analyses for heart-related subjects.

For the nerve theme there are 19 related subjects classified into four association strength tiers. These findings show that 40.43% of all subjects in the subject directory are related to the nerve theme to some extent.

The heart theme includes 13 related subjects grouped into four tiers. About 28% of all the subjects in the subject directory are related to the heart theme to some degree.

It can be seen that although the colour distributions in the six component planes are quite different, the light areas did help the authors identify the related subjects to a specified subject. This study shows that a relatively large SOM display space would produce a better component

plane for association analysis than a relatively small SOM display space, and thus confirms the findings of Ultsch (1995).

While both the SOM method and component plane analysis method can be used for object clustering analysis, they work in different ways. In the SOM method, the objects projected onto the same cell or adjacent cells are regarded as the related subjects, while in the component plane analysis method the objects mapped onto the cells with light colours (high cell values) are regarded as related subjects. This implies that the SOM method cannot identify the relevant objects in its visual space if the objects are not located in the same cell or neighbouring cells, but the component plane analysis method can do so. For instance, in figure 2, children's health (8) is situated separately from brain, nervous system (6), mental health (29), and neurological disorders (31), but it is determined to be the strongest association to the nerve theme in table 2. The same is true with nutrition and herbs (32). Although it was located some distance from cholesterol (9), heart disease (21), and high blood pressure (22) in figure 2, nutrition and herbs (32) was identified as the strongest association to the heart theme in table 2.

Both users' query terms and traversal data on a subject directory can be extracted from a transaction log for users' information-seeking behaviour studies. Subjects in the health-related directory are not like query terms from a transaction log. Query terms are more specific and dynamic than subjects on a directory. Furthermore, query terms come directly from users. This suggests that it is more suitable to make a comparison analysis between the query terms and an established medical thesaurus like *MeSH*.

Conclusion

In this study, a one-year transaction log from the health portal HealthLink was employed. Subjects from its subject directory were selected, and users' traversal information on the subjects was extracted. The SOM technique was applied to a study of health themes and their related subjects because the SOM approach is known to reveal object relationships in visual space (also known as feature map). Two focus themes (heart and nerve) were identified and selected to create the SOM display. Each theme corresponds to a group of related subjects. A component plane analysis was conducted for each subject in the theme group to

[34.228.7.237] Project MUSE (2024-03-19 02:53 GMT)

gain more detailed information about the relationships among the involved subjects. The association judgement of a subject group to a specified subject was made on the basis of their colours in the component plane. Then association analysis results for a theme were merged and summarized. Finally all related subjects to a selected theme were categorized into association strength tiers based on their associations to the theme.

The evaluation results show that results for both the heart theme group and nerve theme group are sound and reasonable. The evaluation results demonstrate that the associations among the related subjects revealed from the visual component analysis method based on user's browsing activities on a health-related directory correspond with the expert's association judgements.

This study shows that users' navigating activities on the health-related directory can be used to reveal and identify the associations among the subjects in the directory, and the component analysis is reliable when it is applied to identifying relationships among subjects in such a situation.

Medical information plays an extremely important role for medical professionals and public health consumers. It is widely recognized that medical terminology is specialized, and usage of medical terminology by medical professionals is quite different from that of health consumers. This study on medical subject associations from the consumers' perspective offers a unique way to understand medical subjects and their relationships. A Web transaction log usually includes both users' query data and navigation data, and both can be used for medical term analysis. This investigation on the subjects from a directory based on users' navigational activities in a Web transaction log provides researchers with a new method to understand semantic relationships among medical subjects, integrating the visual SOM analysis method and the component plane analysis method. The component plane analysis method can identify related objects that are not located in the same cell or adjacent cells in the visual space. These objects cannot be identified in the SOM analysis method that usually identifies the related objects in the same cell and/or adjacent cells in the visual space. This unique feature of the component plane analysis method complements the SOM analysis method. The findings of this study can be used (1) to optimize the medical sub-

ject directory structure by suggesting more related subjects; (2) to revise medical thesauri, which can be enriched by adding related subjects; and (3) to make the content organization and use of a medical portal such as HealthLink more effective and efficient. The study provides not only the subjects related to a subject but also the degree to which they are related to that subject, thus assisting system designers to arrange related subjects or articles more effectively. More closely related subjects or articles are placed prior to the less related subjects or articles in a portal.

Because there is no direct link between two subject nodes in the HealthLink subject directory, the component plane analysis in this study had to be conducted on the basis of browsing activities for path length two. If it were possible to conduct the component plane analysis on the basis of the navigating activities for path length one, the results would have been more reliable. That is one of the limitations of this study. Since this study demonstrates that the presented research method is applicable in the traversal analysis of a transaction log, future research can extend this study to include more subject themes. As a result, the relationships among the subject themes at a higher level can be identified and the corresponding patterns can be discovered.

Notes

1. The HealthLink site was discontinued in May 2009.
2. Figures 2–8 in this article can be viewed in full colour by visiting http://www.utpjournals.com/cjils/cjils.html and following the Supplementary Material link on the right-hand side of the page.
3. If a user browses node A, then node B, the path length from node A to node C is one. Similarly, if a user browses node A, then node B, and finally node C, the path length from node A to node C is two.

References

Abonyi, J., S. Nemeth, C. Vincze, and P. Arva. 2003. Process analysis and product quality estimation by self-organizing maps with an application to polyethylene production. *Computers in Industry* 52 (3): 221–34.

Adami, G., P. Avesani, and D. Sona. 2003. Bootstrapping for hierarchical document classification. Proceedings of the Twelfth International Conference on Information and Knowledge Management (295–302). New Orleans: ACM.

Alexopoulou, D., T. Wächter, L. Pickersgill, C. Eyre, and M. Schroeder. 2008. Terminologies for text-mining: An experiment in the lipoprotein metabolism domain. *BMC Bioinformatics* 9 (Suppl 4). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2367629/.

Begley, C.E., J.F. Annegers, L.B. Lairson, and T.F. Reynolds. 1998. Epilepsy incidence, prognosis, and use of medical care in Houston, Texas, and Rochester, Minnesota. *Epilepsia*, 39 (Suppl 6): S222.

Brooks, T.A. 1995. People, words, & perceptions: A phenomenological investigation of textuality. *Journal of the American Society for Information Science* 46 (2): 103–15.

Calle, E.E., and M.J. Thun. 2004. Obesity and cancer. *Oncogene* 23:6365–78.

Coiera, E.W., and V. Vickland. 2008. Is relevance relevant? User relevance ratings may not predict the impact of Internet search on decision outcomes. *Journal of the American Medical Informatics Association* 15 (4): 542–5.

Díaz, I., A. Cuadrado, A. Diez, L.R. Loredo, F.O. Carrera, and J.A. Rodríguez. 2003. Visual predictive maintenance tool based on SOM projection techniques. *Revue de Métallurgie* 103 (3): 307–15.

Ersoya, O., E. Aydara, A. Gourgaudb, H. Artunerc, and H. Bayhan. 2007. Clustering of volcanic ash arising from different fragmentation mechanisms using Kohonen self-organizing maps. *Computers & Geosciences* 33 (6): 821–8.

Fisk, J.M., P. Mutalik, F.W. Levin, J. Erdos, C. Taylor, and P. Nadkarni. 2003. Integrating query of relational and textual data in clinical databases: A case study. *Journal of the American Medical Informatics Association* 10 (1): 21–38.

Jelier, R., M.J. Schuemie, A. Veldhoven, L.C.J. Dorssers, G. Jenster, and J.A. Kors. 2008. Anni 2.0: A multipurpose text-mining tool for the life sciences. *Genome Biology* 9. http://genomebiology.com/2008/9/6/R96.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biology Cybernetics* 43 (1): 59–69.

Kohonen, T. 1995. *Self-organizing maps*. Berlin: Springer-Verlag.

Kruka, A. , S. Lekb, Y.S. Parkc, and T. Penczak. 2007. Fish assemblages in the large lowland Narew River system (Poland): Application of the self-organizing map algorithm. *Ecological Modeling* 203 (1–2): 45–61.

Laine, J. 1998. Analysis and monitoring of continuous casting mould with the self-organizing map. Proceedings of the ECSC Workshop on Application of Artificial Neural Network Systems in the Steel Industry, Brussels, January 1988, 151–5.

Lin, J. 2008. PageRank without hyperlinks: Reranking with PubMed related article networks for biomedical text retrieval. *BMC Bioinformatics* 9: 270.

Liu, H., and C. Friedman. 2003. Mining terminological knowledge in large biomedical corpora. *Pacific Symposium on Biocomputing* 8: 415–26.

Mohanty, S.K., A.T. Mistry, W. Amin, A.V. Parwani, A.K. Pople, L. Schmandt, S.B. Winters, E. Milliken, P. Kim, N.B. Whelan, G. Farhat, J. Melamed, E. Taioli, R. Dhir, H.I. Pass, and M.J. Becich. 2008. The development and deployment of Common Data Elements for tissue banks for translational research in cancer: An emerging standard based approach for the Mesothelioma Virtual Tissue Bank. *BMC Cancer* 8: 91.

Pei, J., J. Han, B. Mortazavi-asl, and H. Zhu. 2000. Mining access patterns efficiently from Web logs. Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto, Japan, 18–20 April, 396–407.

Pesquita, C., D. Faria, H. Bastos, A.E.N. Ferreira, A.O. Falcão, and F.M. Couto. 2008. Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics* 9 (Suppl 5): S4.

Postolache, O.A., P.M.B.S. Giro, J.M.D. Pereira, and H.M.G. Ramos. 2005. Self-organizing maps application in a remote water quality monitoring system. *IEEE Transactions on Instrumentation and Measurement* 54 (1): 322–9.

Pratt, W., and L. Fagan. 2000. The usefulness of dynamically categorizing search results. *Journal of American Medical Information Association* 7 (6): 605–17.

Tsuruoka, Y., J. McNaught, and S. Ananiadou. 2008. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics* 9 (Suppl 3). http://www.biomedcentral.com/1471-2105/9/S3/S2.

Tsuruoka, Y., J. McNaught, J. Tsujii, and S. Ananiadou. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics* 23 (20): 2768–74.

Ultsch, A. 1995. Self-organizing neural networks perform different from statistical k-means clustering. Proceedings of Gesellschaft für Klassifikation, Basel, Switzerland, March.

Wang, H., M. Huang, S. Ding, and X. Zhu. 2008. Exploiting and integrating rich features for biological literature classification. *BMC Bioinformatics* 9 (Suppl. 3) http://www.biomedcentral.com/1471-2105/9/S3/S4.

Zhang, J., L. An, T. Tang, and Y. Hong. 2009. Visual health subject directory analysis based on users' traversal activities. *Journal of the American Society for Information Science and Technology* 60 (10): 1977–94.

Zhang, J., D. Wolfram, P. Wang, Y. Hong, and R. Gillis. 2008. Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science and Technology* 59 (11): 1–15.

Zhang, J., and D. Wolfram. 2009. Obesity-related query term analysis in a public health portal transaction log. *Online Information Review* 33 (1): 43–57.

Zhou, B.Y., S.C. Hui, and A.C.M. Fong. 2006. Efficient sequential access pattern mining for Web recommendations. *International Journal of Knowledge-Based and Intelligent Engineering Systems* 10 (2): 155–68.