Visual analysis of obesity-related query terms on HealthLink

Jin Zhang and Dietmar Wolfram
School of Information Studies, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Abstract
Purpose – The purpose of this article is to investigate obesity-related queries from a public health portal (HealthLink) transaction log.
Design/methodology/approach – Multidimensional scaling (MDS) was applied to each of five obesity-related focus keywords and their co-occurring terms in submitted queries. After the transaction log data were collected and cleaned, and query terms were extracted and parsed, relationships between a focus keyword and its co-occurring terms were established. Clustering relationships between focus keywords and their co-occurring terms were identified and analysed in the MDS visual context.
Findings – The MDS analysis produced satisfactory outcomes for all five focus keywords. The term “placements”, in the visual configurations revealed strong grouping tendencies of three to five clusters for each focus keyword.
Originality/value – The findings of this study provide insights into health consumers’ internet-based information-seeking behaviour on obesity-related topics. These findings could be used to enhance online search system design and health-related thesaurus construction.

Keywords Obesity, Information systems, Health services, United States of America, Information searches, Online catalogues

Paper type Research paper

Introduction
Health informatics is an emerging field. Of growing interest is the information behaviour of health information consumers. It is widely recognised that the search behaviours of medical professionals like doctors and physicians are quite different from those of health consumers in terms of search vocabulary (Ratzan and Parker, 2000; Zeng and Tse, 2006). A better understanding of health consumer search behaviours would benefit both medical professionals (by allowing them to communicate better with patients) and health consumers (by enabling them to express their health needs more effectively).

An example of a popular search topic where a better understanding of health consumer behaviour is needed is the area of obesity. Obesity is an epidemic in not only North America but also in other parts of the world. It has become a major health concern that can lead to serious health problems such as cardiovascular disease, diabetes mellitus type 2, sleep apnoea and osteoarthritis (Haslam and James, 2005). Obesity currently results in an estimated 400,000 deaths annually in the USA (Mokdad...
et al., 2000) and costs nearly US$122.9 billion (National Institutes of Health: National Institute of Diabetes, Digestive and Kidney Diseases, 2008). It is reported that obesity is the second leading cause of preventable death in the USA (American Obesity Association, 2004). Obesity not only has a negative impact on lifestyle but can also lead to lower self-esteem, depression and discomfort in social situations (The Hormone Foundation, 2004). It is not surprising that obesity is increasingly regarded as one of the most serious and growing public health problems.

The information behaviour of health consumers can be studied using different means, including interviews, surveys and transactions collected from health information systems. Transaction log analysis permits the study of electronically recorded interactions between online information retrieval systems and the persons who search for the information found in those systems (Peters, 1993). The ability to conduct transaction log analysis has been further expanded and enriched thanks to the internet. One advantage of a transaction log approach to the study of health information-seeking behaviour is that a web log file accurately records all online users’ activities performed on a server. Usually, the log of a user’s request includes a client IP address, request date/time, page requested, HTTP code, number of bytes transferred, user agent, referrer, etc. Embedded in these logs are also the queries users have submitted. Unlike interview approaches to collecting user data, transaction logs can amass large quantities of data for analysis.

Transaction log analysis consists of the following steps: data collection, data cleaning, user identification, session identification, feature selection, data transformation, data combination, mining the data and results visualisation (Pabarskaite and Raudys, 2007). With the trend of system analysis and design shifting from a system-centred to a user-centred perspective, transaction log analysis plays a more important role in improving information organisation, information presentation and the functionality of an information system, which can enhance the information-seeking efficiency of searchers and provide a better understanding of users’ search behaviour patterns.

There have been many examples of studies since the mid-1990s that have relied on transaction logs to highlight user information seeking. Studies of query characteristics have included special topic web sites such as THOMAS (Croft et al., 1995), academic web sites (Wang et al., 2003), digital libraries (Jones et al., 1998), web-based OPACs (Cooper, 2001) and bibliographic databases (Yi et al., 2006), as well as public search engines such as Alta Vista (Silverstein et al., 1999), Excite (Jansen et al., 2000) and Vivisimo (Koshman et al., 2006), and federated search systems such as Dogpile (Jansen et al., 2007).

Analytical techniques have been applied to transaction logs to identify patterns in the log files. For example, Perkowitz and Etzioni (1997) used artificial intelligence techniques to automatically expand their portal administration and management by learning from user access patterns to create adaptive web sites. Web log analysis enables reorganisation and optimisation of a web site to facilitate searchers’ access to the desired pages (Pirolli et al., 1996). Transaction log analysis can also help users with online searching. After related queries in a web log are identified and modified by an enhanced model and association rules, the results can be used to assist users with initial search query input (Shi and Yang, 2007; Huang et al., 2003).

Web log clustering analysis, which can reveal hidden structures and patterns among an investigated dataset, has attracted researchers’ attention for some time. In a
cluster analysis, searchers or visitors are treated as objects and their browsing behaviour and characteristics categorised into distinctive groups (Hand et al., 2001; Jain et al., 1997; Xiao and Zhang, 2001). Beitzel et al. (2007) investigated properties of a very large query log over varying periods, and were able to identify and illustrate topical trends in a temporal context.

One way of better understanding search behaviours in this area is through information visualisation, which utilises human perception and extends human cognitive capacity to comprehend information. Information visualisation is particularly powerful for revealing the connections and relationships among investigated objects. For this reason, information visualisation can be used to support tasks such as data analysis, information exploration, information explanation, trend prediction and pattern detection (Zhang, 2008). Visualisation is not only a unique method for information presentation, it is also an effective method for data analysis (Card et al., 1999).

There are many information visualisation techniques and methods available. As one of the most widely used and mature information visualisation techniques, multidimensional scaling (MDS) serves as a useful tool for this purpose (Kruskal, 1964). The MDS approach has been applied to document co-citation analysis (York et al., 1995), journal co-citation analysis (Hakanen and Wolfram, 1995), subject co-citation analysis (Small and Garfield, 1985) and webpage co-citation analysis (Thelwall, 2002; Vaughan, 2006).

The purpose of the study was to determine whether the analysis of query terms related to obesity can reveal patterns and health consumer information needs using MDS as an information visualisation tool. Such knowledge can enhance our understanding of health consumer information behaviour and provide insights into health consumer vocabulary usage, which can benefit information system design and controlled vocabulary construction.

**Research methodology**

Web transaction logs allow researchers to trace first-hand, detailed, authentic and complete information of users. These characteristics make it possible to better understand web site user behaviours, and identify patterns and potential problems associated with system usage.

As this study focused on the topic of obesity, a group of focus keywords with a close semantic relationship to obesity were identified using Medical Subject Headings (MeSH, 2008), a well-known medical thesaurus that consists of sets of terms naming descriptors in a hierarchical structure. Five focus keywords relevant to obesity were identified from the MeSH annotation, scope note, related terms and allowable qualifiers. These keywords were:

1. obesity;
2. fat;
3. diet;
4. food; and
5. weight.

Visual analysis of query terms
The keywords were then extracted from queries in the transaction log and their co-occurring query terms were identified and analysed. A visual configuration in the MDS context was generated for analysis for each focus keyword and the frequent terms with which it co-occurred in queries. The distributions of terms within each visual configuration were characterised.

The investigated dataset was a one-year (2005) user log file collected from HealthLink (2008), a public health consumer portal developed by the Medical College of Wisconsin. HealthLink’s goals are to promote the health and well-being of the community and to provide health consumers with accurate, timely, impartial and authoritative medical information. During the data collection period, the log recorded more than 370,000 queries and other information.

A transaction log contains rich information ranging from access time and bytes transferred to queries submitted by users. For this study, the analysis focused on query terms. The raw transaction log was first cleaned to exclude useless data and to preserve the useful data for further processing in a Microsoft Access database. Queries were parsed for identifiable terms, consisting of strings of alpha-numeric characters, usually delimited with non-alpha-numeric characters. Terms and their raw frequencies in the transaction log were tallied to form a term master file. Each of the five focus keywords on obesity were checked against the term master file to make sure that the corresponding raw frequency in the transaction log was large enough for a term co-occurrence analysis. Generally speaking, the higher the frequency of a term in the term master file, the more co-occurring terms it is likely to have in the transaction log. If a focus keyword co-occurred with a very small number of terms, it would be difficult to generate a sound and meaningful proximity analysis.

To conduct an MDS analysis for a focus keyword and the terms with which it co-occurred, all co-occurring terms needed to be identified. It is clear that not all co-occurring terms would be appropriate for inclusion in the analysis because low frequency terms and some meaningless terms with high frequencies would make little or no contribution to the analysis. Therefore, those terms with low co-occurrences were excluded and terms that served a grammatical function or provided no subject context were filtered out.

In the MDS visual analysis, a focus keyword and its co-occurring terms were projected onto a three-dimensional visual space. The resulting visual configuration of all co-occurring terms illustrated semantic relationships among the focus keywords. Terms that are semantically connected are more likely to be clustered together in the visual space, whereas terms that are not semantically related are more likely to be further apart. In order to establish the visual configuration for the terms, the similarity between two terms had to be defined. The similarity (proximity) relationships among the terms were described and presented in an $n \times n$ square matrix ($n$ is the number of the involved terms). The similarity/proximity is defined in equation (1).

$$Sim(t_i, t_j) = \frac{F_{ij}}{MIN(F_i, F_j)} \quad (1)$$

where $F_i$ ($F_j$) is the frequency of term $t_i$ ($t_j$) and $F_{ij}$ is their frequency of co-occurrence.

Relationships among the terms are presented in the $n \times n$ matrix, which represents an $n$-dimensional space. Because of the difficulties associated with interpreting relationships in a high dimensional space, the MDS visual space projects terms onto a
two-dimensional or three-dimensional space, which are more easily interpreted. In other words, the semantic and abstract relationships among the terms in the high dimensional space are transformed into proximity-based relationships in the lower dimensional space. The proximity or distance in the low visual space can be defined by the Minkowski metric (Korfhage, 1997).

\[ d_{ij} = \left( \sum_{r=1}^{n} (a_r - b_r)^k \right)^{\frac{1}{k}}, k = 1, \ldots, \infty \]  

In equation (2), given \( x_i = (a_1, a_2, \ldots a_n) \) and \( x_j = (b_1, b_2, \ldots b_n) \), which represent two terms in a high dimensional space, \( n \) is the dimensionality of the space. In fact, \( x_i = (a_1, a_2, \ldots a_n) \) and \( x_j = (b_1, b_2, \ldots b_n) \) represent two corresponding rows in the proximity matrix, respectively. Therefore, the distance between \( x_i = (a_1, a_2, \ldots a_n) \) and \( x_j = (b_1, b_2, \ldots b_n) \) in the low dimensional visual space is defined as \( d_{ij} \). The Minkowski parameter \( k \) can define a family of the Minkowski metric. In this study the parameter \( k \) was set to either 1 or 2.

Due to the dimensionality reduction, the relationships among the keywords in the high dimensional space may not be faithfully demonstrated in the low dimensional visual space. Some relationship distortion is inevitable. The solution is to find a way to minimise the distortion after the projection. A mechanism to measure the distortion, called the stress value (S), is defined in equation (3).

\[ S = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (f(\delta_{ij}) - d_{ij})^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} (d_{ij})^2}^{1/2} \]  

Here \( n \) is equal to the number of all terms involved, \( d_{ij} \) is the Euclidean distance between two terms \( x_i \) and \( x_j \) in the low visual space and \( f(\delta_{ij}) \) is the similarity/proximity between terms \( x_i \) and \( x_j \) in the high dimensional space.

The last issue is the dimensionality of the low MDS visual space. In order to observe projected terms in the visual space, the dimensionality must be no larger than three. There is a long-standing debate over two-dimensional presentation versus three-dimensional presentations in the information visualisation community. Without a doubt, each presentation method has its own strengths and weaknesses in terms of technical implementation, object control, richness of information presentation, nature of visualised objects, etc. In addition, the dimensionality of the low visual space affects the stress value of the MDS results. Considering these factors, a three-dimensional visual space was used for this study to achieve better object control and a better stress value for the MDS results.

SPSS ALSCAL was used to perform the MDS analysis due to the software’s robustness and three-dimensional MDS display platform. It is worth pointing out that SPSS generates the squared correlation index RSQ (\( R^2 \)) for the MDS analysis result. It is also designed to evaluate the quality of an MDS result. The lower the stress value and the higher the \( R^2 \) value, the better the MDS result is.
Results and analysis
After a focus keyword and its co-occurring terms were singled out in the transaction log, irrelevant terms were discarded and proximity relationships among these terms were defined and normalised. Data were then entered into SPSS for visual analysis. A summary of the results for the five identified focus keywords (obesity, fat, diet, food and weight), their raw term frequencies, the corresponding Minkowski parameters, resulting RSQ ($R^2$), stress values, numbers of displayed keywords in the visual space, and the numbers of identified groups or clusters observed for each of the focus keywords are summarised and listed in Table I.

The focus keyword raw frequencies ranged from 336 (fat) to 3462 (diet), and the average raw frequency was 1687.8. The Minkowski parameters were set to either 1 or 2 for all five focus keywords. All values of RSQ ($R^2$) were larger than 0.98, and the average RSQ ($R^2$) was 0.989894. All stress values were smaller than 0.090, with an average stress value of 0.078774. Note that stress values below 0.10 and $R^2$ values above 0.90 are considered sound and satisfactory, indicating that each outcome is acceptable for interpretation. The number of terms included in each analysis ranged from 18 (fat) to 53 (diet), with an average of 36.2 terms. Between three and five clusters of terms for each focus keyword were inferred from the MDS visual output.

It should be noted that because a focus keyword was related (relevant) to all its co-occurring terms in the visual space, the focus keyword was not grouped into any clusters in the visual analysis. Terms are abbreviated to $Vx$ (where $x$ is a number) in each figure, with the term to which each abbreviation corresponds appearing immediately below each figure. Space limitations and improved visual aesthetics warranted the use of the abbreviations.

**Obesity**
Figure 1 contains the result for the focus keyword obesity ($V22$), from which five clusters emerged. In cluster C3, three terms (childhood, children and kids) suggest that childhood obesity is an important concern for searchers. Terms in cluster C5 such as beverage, environmental and gene reflect topics related to causes of obesity.

C1: Genes, human, responsible, zeroing.
C2: Health, problems, related, severe.
C3: Address, blood, factors, general, pressure, surgeon, US.
C4: Cause, causes, childhood, children, diabetes, gastric, genetic, joint, kids, pain, surgery, test.

<table>
<thead>
<tr>
<th>Focus keyword</th>
<th>Keyword raw frequency</th>
<th>Minkowski parameter</th>
<th>RSQ ($R^2$)</th>
<th>Stress value</th>
<th>Number of terms</th>
<th>Number of clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obesity</td>
<td>2,205</td>
<td>1</td>
<td>0.98892</td>
<td>0.09210</td>
<td>38</td>
<td>5</td>
</tr>
<tr>
<td>Diet</td>
<td>3,462</td>
<td>2</td>
<td>0.99310</td>
<td>0.07458</td>
<td>53</td>
<td>4</td>
</tr>
<tr>
<td>Weight</td>
<td>1,441</td>
<td>2</td>
<td>0.98931</td>
<td>0.08846</td>
<td>37</td>
<td>4</td>
</tr>
<tr>
<td>Food</td>
<td>995</td>
<td>1</td>
<td>0.98979</td>
<td>0.08444</td>
<td>35</td>
<td>3</td>
</tr>
<tr>
<td>Fat</td>
<td>336</td>
<td>2</td>
<td>0.98835</td>
<td>0.06629</td>
<td>18</td>
<td>3</td>
</tr>
</tbody>
</table>

Table I.
Data analysis summary of the five focus keywords
**Query terms:** address (V0), associated (V1), beverages (V2), blood (V3), cause (V4), causes (V5), childhood (V6), children (V7), contribute (V8), diabetes (V9), environmental (V10), factors (V11), gastric (V12), gene (V13), general (V14), genes (V15), genetic (V16), health (V17), high (V18), human (V19), joint (V20), kids (V21), obesity (V22), pain (V23), pressure (V24), problems (V25), related (V26), research (V27), responsible (V28), scientific (V29), severe (V30), surgeon (V31), surgery (V32), term (V33), test (V34), time (V35), US (V36), zeroing (V37)

C5: Associated, beverages, contribute, environmental, gene, high, research, scientific, term, time.

**Diet**

Figure 2 shows the result for the focus keyword diet (V17). The four visible clusters represent three defined topic areas and a larger more generic group. In cluster C2 American, Heart and Association were grouped together to reflect searches about the American Heart Association. The terms blood and clots appeared in the same cluster (C3). Although the largest cluster (C4) was somewhat broad in its content, sub-themes are apparent, where calorie, eat, protein, sodium, fiber and foods relate to consumption; guidelines, health, healthy, prevent, plan, exercise and balanced relate to health issues; and people, adult and children refer to different groups.

C1: Patient, syndrome, nephritic, body, effects, human.

C2: Association, American, heart, low.

C3: Blood, clots, lungs.

Figure 2.
Display of focus keyword diet

Query terms: adult (V0), American (V1), association (V2), balanced (V3), blood (V4), body (V5), breast (V6), calorie (V7), cancer (V8), cardiac (V9), chart (V10), children (V11), cholesterol (V12), chronic (V14), clots (V15), diabetic (V16), diet (V17), disease (V18), diverticulitis (V19), diverticulosis (V20), eat (V21), effects (V22), exercise (V23), fat (V24), fiber (V25), food (V26), foods (V27), guidelines (V28), health (V29), healthy (V30), heart (V31), high (V32), human (V33), kidney (V34), low (V35), lower (V36), lowering (V37), lungs (V38), menu (V39), nephritic (V40), patient (V41), people (V42), plan (V43), potassium (V44), pregnancy (V45), pressure (V46), prevent (V47), protein (V48), sarcoidosis (V49), sodium (V50), surgery (V51), syndrome (V52), treatment (V53)

Weight

Figure 3 summarises the result for the focus keyword weight (V35), which contains four identifiable clusters. Cluster C1 groups together terms related to high blood pressure. Cluster C2 groups terms associated with physical activity and, oddly enough, bodily functions. Cluster C4 groups terms related to medical issues for different groups (children, male and women).

C1: Blood, high, pressure.

C2: Activity, control, physical, bowel, movement, regular, limits.

C3: Syndrome, loss, dumping.

C4: Age, backpack, body, breast, cancer, cause, children, diabetes, diet, fat, gain, height, ideal, infection, lose, male, management, medication, program, rapid, reduction, sudden, women.

Food

Figure 4 contains the result for the focus keyword food (V15), which produced three identifiable clusters. Cluster C1 reflects concerns related to Americans’ ingestion of large portion sizes. Cluster C2 contains two themes related to vitamins C and K, as well
Visual analysis of query terms

Figure 3. Display of focus keyword weight

Query terms: activity (V0), age (V1), backpack (V2), blood (V3), body (V4), bowel (V5), breast (V6), cancer (V7), cause (V8), children (V9), control (V10), diabetes (V11), diet (V12), dumping (V13), fat (V14), gain (V15), height (V16), high (V17), ideal (V18), infection (V19), limits (V20), lose (V21), loss (V22), male (V23), management (V24), medication (V25), movement (V26), physical (V27), pressure (V28), program (V29), rapid (V30), reduction (V31), regular (V32), sudden (V33), syndrome (V34), weight (V35), women (V36)

Figure 4. Display of focus keyword food

Query terms: Americans (V0), avoid (V1), blood (V2), borne (V3), C (V4), calcium (V5), cholesterol (V6), clostridial (V7), diabetic (V8), diet (V9), disease (V10), diverticulitis (V11), eat (V12), fast (V13), fiber (V14), food (V15), guide (V16), high (V17), ignore (V18), importance (V19), K (V20), list (V21), long (V22), patient (V23), poisoning (V24), portion (V25), portions (V26), pressure (V27), rich (V28), safety (V29), size (V30), sources (V31), symptoms (V32), types (V33), vitamin (V34), water (V35)
as clostridium food poisoning. Cluster C3, again representing more of a mixed bag of themes, reflects themes of high blood pressure and diseases.

C1: Ignore, importance, portions, Americans, portion, size.
C2: Poisoning, clostridial, types, K, sources, vitamin, C.
C3: Avoid, blood, borne, calcium, cholesterol, diabetic, diet, disease, diverticulitis, eat, fast, fiber, guide, high, list, long, patient, pressure, rich, safety, symptoms, water.

Fat
Figure 5 shows the result for the last focus keyword fat (V7). The three clusters represent male issues (C1), dietary concerns (C2) and recommended dietary intake (C3).

C1: Body, male, reduce.
C2: Diet, diets, low, cholesterol, foods, content, sodium, tables.
C3: Breast, daily, grams, intake, recommended, saturated, total.

Although each focus keyword was associated with a group of co-occurring terms, which varied among the focus keywords, these terms can be categorised into relevant professional associations, treatments for a disease, causes and reasons for a disease, prevention of a disease, tests for a disease, symptoms of a disease, etc.

It is no surprise that the selected focus keywords shared a number of co-occurring terms because the focus keywords are closely related. Table II outlines all the focus keywords and the distribution of their shared co-occurring terms. A total of 26 terms
were shared by at least two focus keywords. The most frequently shared terms were diet(s) (4), blood (4), high (4) and pressure (4). This implies that high blood pressure and diet(s) are major themes associated with obesity for online health searchers. Other frequently shared terms shared by three focus keywords were: cholesterol, low, body, food, breast, children and fat.

The MDS analysis of terms associated with obesity revealed themes in searches on this topic that would not be evident from listing query terms on their own. Issues associated with health, groups of people (male, female and children), specific diseases and conditions, and dietary habits are evident based on the proximities of search terms to one another in conjunction with the focus keywords.

The dataset and the analysis technique used provide a portrait of health consumer information searching. However, as with other similar studies, there are limitations to the use of transaction logs. For example, the queries themselves cannot tell investigators why searchers searched for the topics they did or why they selected the terms they did. But even the queries by themselves can be revealing of search behaviour.

Unlike other transaction log studies for which datasets may have been collected over a limited period of time, the HealthLink dataset represented one year of queries, which minimised the occurrence of any seasonal variations in search topics.

<table>
<thead>
<tr>
<th>Term</th>
<th>Obesity</th>
<th>Diet</th>
<th>Weight</th>
<th>Fat</th>
<th>Food</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diet(s)</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>4</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>Low</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Body</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Food</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Sodium</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Breast</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>American(s)</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Blood</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Diabetic</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Disease</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Diverticulitis</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Eat</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Fiber</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Patient</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Pressure</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Syndrome</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Cancer</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Cause(s)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Children</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Fat</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>Human</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Health(y)</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Surgery</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>24</td>
<td>12</td>
<td>8</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>
MDS serves as an exploratory tool and provides outcomes that must be interpreted, including the term groupings. Outcomes do not provide indisputable proof on which conclusions may be drawn. Also, multiple explanations for term groupings may exist. However, given the excellent stress and RSQ values associated with the outcomes, along with the well-defined clustering of terms, the identified groups are based on strong empirical evidence. The interpretation of the observed clusters, in any case, will be subjective.

**Conclusion**

It is widely recognised that an obesity health crisis is looming. Obesity increases the risk of diabetes, heart disease, stroke, arthritis and other diseases. Scientists and medical professionals are looking into the many biological factors underlying obesity. But how the public responds to the obesity epidemic is a mystery. This study sheds light on the issue from this unique angle.

In this study, a public health information portal transaction log file was examined to gain insight into the information-seeking behaviour of internet searchers on the topic of obesity. Based on MeSH subject headings in conjunction with term raw frequencies in the master query term file, five focus keywords (obesity, fat, diet, food and weight) were selected for the study.

The original HealthLink transaction log data was cleaned, useful query terms were separated, less frequent and useless words were excluded, focus keywords and their related keywords were singled out, and a proximity relationship between the involved keywords was established. The MDS visual analysis method was applied to each of the five focus keywords and their co-occurring terms for a visual cluster analysis. The positive MDS stress values and RSQ values of the investigated focus keywords indicate that each of the MDS visual analyses yielded satisfactory and sound results.

The authors found that topics of diet and high blood pressure were primary concerns associated with obesity in searches by health consumers, based on their frequency of appearance as co-occurring terms among the focus keywords. For each focus keyword, three to five co-occurring term clusters were produced. Terms in an identified cluster are semantically associated.

If an obesity-related keyword is used in a query, its co-occurring terms in the cluster can be used to either expand or revise the query. The findings of the study can also be used to revise health-oriented thesauri by adding more end-user-centred terminology or vocabulary. For instance, for a focus keyword like diet, the related terms for this keyword in existing controlled vocabularies and thesauri, such as MeSH or SNOMED CT (Systematised Nomenclature of Medicine – Clinical Terms), can be compared with the co-occurring terms from the transaction log associated with diet. Frequently co-occurring terms from the logs that do not appear in the thesauri may warrant inclusion as related terms in the thesauri. Finally, the findings can help health and medical professionals to better understand patient terminology or vocabulary related to obesity.

**References**


Jain, N., Han, E., Mobasher, B. and Srivastava, J. (1997), Web Mining: Pattern Discovery from World Wide Web Transactions, University of Minnesota, Minneapolis, MN.


About the authors

Jin Zhang is Associate Professor in the School of Information Studies, University of Wisconsin-Milwaukee. He received his PhD from the University of Pittsburgh. His research interests include visualisation for information retrieval, internet information organisation, data mining and information retrieval. Jin Zhang is the corresponding author and can be contacted at: zhang@sois.uwm.edu

Dietmar Wolfram is also a Professor in the School of Information Studies, University of Wisconsin-Milwaukee. He received his PhD from the University of Western Ontario. His research interests include information retrieval system design and evaluation, as well as applied informetrics.