Information Processing and Management 49 (2013) 1019–1048

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A user term visualization analysis based on a social question and answer log

Jin Zhang [a,*], Yiming Zhao [b,1]

[a] School of Information Studies, University of Wisconsin Milwaukee, Milwaukee, WI 53201, United States
[b] Center for Studies of Information Resources, Wuhan University, Wuhan, Hubei Province 430072, PR China

ARTICLE INFO

Article history:
Received 2 October 2012
Received in revised form 18 April 2013
Accepted 26 April 2013

Keywords:
Diabetes
Term analysis
Information seeking pattern
Visualization analysis
Social question and answers
Q&A

ABSTRACT

The authors of this paper investigate terms of consumers' diabetes based on a log from the Yahoo!Answers social question and answers (Q&A) forum, ascertain characteristics and relationships among terms related to diabetes from the consumers' perspective, and reveal users' diabetes information seeking patterns. In this study, the log analysis method, data coding method, and visualization multiple-dimensional scaling analysis method were used for analysis. The visual analyses were conducted at two levels: terms analysis within a category and category analysis among the categories in the schema. The findings show that the average number of words per question was 128.63, the average number of sentences per question was 8.23, the average number of words per response was 254.83, and the average number of sentences per response was 16.01. There were 12 categories (Cause & Pathophysiology, Sign & Symptom, Diagnosis & Test, Organ & Body Part, Complication & Related Disease, Medication, Treatment, Education & Info Resource, Affect, Social & Culture, Lifestyle, and Nutrient) in the diabetes related schema which emerged from the data coding analysis. The analyses at the two levels show that terms and categories were clustered and patterns were revealed. Future research directions are also included.

1. Introduction

Diabetes is a chronic disease which causes high levels of sugar in the human blood. Diabetes can result in many complications such as heart disease and stroke, obesity, high blood pressure, blindness, kidney disease, and nervous system related diseases. According to a statistic report of American Diabetes Association (Diabetes statistics, 2012), 25.8 million people in the United States (8.3% of the population) have diabetes, diabetes cost totaled $174 billion in the United States in 2007, and 346 million people worldwide have diabetes (World Health Organization (WHO), 2012). It is not surprising that diabetes is listed as number 7 on the list of 10 leading causes of death in the US 2009 (Center for Disease Control and Prevention (CDC), 2012).

With a common and typical chronic disease, diabetes patients need to become their own managers and personal experts of their treatment (Nordfeldt, Johansson, Carlsson, & Hammersjo, 2005). The treatment of diabetes consists of self-care to be performed by the patients and their families (Brink, Miller, & Moltz, 2002). The studies suggest that the patients search for information about treatment, medicine, diet, and support related to diabetes. It is recognized, however, that there is a huge gap between health consumers (such as diabetes patients) and health information resources. Researchers have tried to bridge the gap (Milewski & Chen, 2010) and in one such study it was reported that there is a significant difference when

* Corresponding author. Tel.: +1 414 229 2712.
E-mail addresses: jzhang@uwm.edu (J. Zhang), zym_0418@qq.com (Y. Zhao).
[1] Tel.: +86 18942901531.

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.04.003

medical professionals and common people use the same medical concepts (Zhang, Wolfram, Wang, Hong, & Gillis, 2008). In order to overcome the obstacle that prevents people from effectively accessing health information, their health information needs on a topic like diabetes should be analyzed, and their information seeking patterns should be identified and studied to empower health consumers to effectively obtain their own health information.

Thanks to the Web 2.0 technology which provides an open user-centered environment for information sharing and information exchanging, patients of diabetes can create an open social Q&A discussion forum on diabetes where they can make contributions to the topic, receive assistance from the forum, and interact with each other in that community. Questions related to the topic can be asked and posted in the forum, and answers to these questions can be provided by other participants who have expertise and experience on the topic. The social Q&A forum has become an important information resource for people to seek information and obtain answers to questions on diabetes.

Information visualization has enormous potential for researchers. Information visualization methods and techniques can be employed as: (1) an effective and intuitive means for people to explore and discover information in a visual environment; (2) an effective research analysis mechanism for researchers to better understand trends in complex data sets within a visual and meaningful display, and reveal hidden and complicated patterns in the data set; and (3) a unique and efficient way to interact with users as well (Zhang, 2008). Due to these reasons, the information visualization method is ideal for studying diabetes information seeking in the social Q&A forum and was used as a primary research method in this study.

The purpose of this study is to use a mixed research method to investigate consumers' diabetes term use patterns based on the log from a social Q&A forum, ascertain characteristics and relationships among terms related to diabetes from the consumers' perspective, reveal term use patterns on diabetes, and discuss how term use patterns can be used to inform thesaurus and classification systems design. The overarching research question is whether consumer's term use patterns on diabetes can be revealed from a social Q&A data set.

Findings of this study can be used to: (1) better understand users' diabetes term use behaviors in terms of the term associations in a social Q&A data set; (2) reveal and demonstrate hidden patterns of users' term use that emerge from Q&A data analysis by the identified and illustrated semantic connections in the themes related to diabetes; (3) update and revise existing diabetes related thesauri, or subject heading systems, or classifications by recommending new related candidate terms emerging from this study to the thesauri, tinkering subject structures, and adjusting relationships between subjects in the subject heading systems and classifications; (4) organize and present diabetes related information in Q&A websites, or other diabetes related health portals, or patient-oriented electric medical information systems by using the emerging subject schema. In fact, the emerging schema as a subject directory can be directly employed to arrange contents of diabetes related websites which guides users to navigate the websites and portals effectively and makes them more user-friendly and user-oriented; and (5) present a new mixed research methodology, which can apply the Q&A data analysis method, coding analysis method, and information visualization method to similar research topics in the health information field.

2. Related research

It is widely recognized that the Internet plays an increasingly important role in users' health information seeking. The Web is an important source for people who are seeking healthcare information (Hesse et al., 2005). Due to the richness and availability of health information on the Internet some people would like to search in there rather than see a doctor (Milewski & Chen, 2010). The Pew Internet Project estimated that at least 75% of US Internet users searched for health information online, with 8 million Americans seeking health information online in a typical day (Fox, 2006, 2008).

Diabetics like other health consumers increasingly use the Internet for their medical consulting and problem solving in addition to traditional information seeking channels such as readings, health literacy, magazine articles, patients' medical records and health care providers. (Brink & Chiarelli, 2004; Milewski & Chen, 2010). Because of the inherent characteristics of diabetes as a long term self-care disease with a lot of available online health information it is no surprise that diabetics rely heavily on the Internet-based information (Nordfeldt et al., 2005). Diabetics seek medical and diabetes-specific information, as well as related information on diet, exercise and stress management, to control their disease (Longo et al., 2010). In addition, millions of non-diabetics, who have a family history of diabetes, who are overweight or who have symptoms similar to diabetes, also worry about their or their friends' health conditions and use online Internet searching for diabetes information.

Studies on diabetics information seeking have been widely reported. Recent research suggests that obtaining health information and staying informed is crucial for diabetics, pre-diabetics and non-diabetics to maintain self-care or to eliminate unnecessarily concerns. Searchers are still struggling with barriers to obtaining such information. Milewski and Chen (2010) concluded that lack of motivation, passiveness, inconsistent information, generality of information, and loss of information are five major barriers for seeking health information. In a study, Longo et al. (2010) identified how individuals with diabetes seek and use health care information. This study used 9 focus groups including 46 adults with diabetes and then analyzed the transcripts and notes from these focus groups. They found diabetics looked for credible sources for their unique needs and information fitting their own knowledge and experiences. Findings in a study (Nordfeldt et al., 2005) on young diabetics' self-education showed that 42% of respondents had searched for diabetes information on the Internet, and 97% anticipated future online searching on the topic. Studies on patients' and parents' attitudes toward an online diabetes portal tailored to young patients with type I diabetes and their parents were investigated and the interactive online diabetes portals

were preferred by young users (Nordfeldt, Hanberger, & Bertero, 2010). Studies on information seeking of diabetics not only benefit common people but also provide useful evidence for medical professionals to understand how diabetics receive or seek information (Longo et al., 2010). In the meanwhile, evaluation and assessment of the quality of diabetes consumer-information Web sites also attracts researchers' attention (Seidman, Steinwachs, & Rubin, 2003).

One popular source of online information is question and answer forums which are both social and interactive. Unlike traditional Q&A services where experts or authorities in a field of interest are responsible for answering questions from users, a social Q&A forum/site forms a community where its members can post questions, answer other members' questions, and rate or rank other members' answers to the questions (Roush, 2006). Questions and answers are usually organized under topical categories which, typically, are archived and searchable (Rosenbaum & Shachaf, 2010). Importance and preference of questions and answers can be ranked by users (Shah, Oh, & Oh, 2008). Not only is information shared in a social Q&A site, but also experience, opinion, and fun, indicating that the function of online communities is both informational and emotional support (Gooden & Winefield, 2007). Social Q&A sites have become some of the most popular destinations for online information seekers (Shah et al., 2008). From 2008 to 2010, the number of visits to the top five Q&A sites has increased by 889% (Rosenbaum & Shachaf, 2010). In particular, Yahoo!Answers has become the most popular Internet reference site in America (Alexa, 2012).

The existing research on social Q&A forums is divided into content-centered studies (questions and answers) and user-centered studies (questioners, answerers, and the community in general) (Shah, Oh, & Oh, 2009). Gazan (2011) summarizes the major threads of emerging social Q&A sites research including user-generated and algorithmic question categorization, answer classification, answer quality assessment, user satisfaction, reward structures, motivation for participation, and trust and expertise.

The astonishing size of the community, the great diversity of information exchanged, and the high quality answers in aggregation make Yahoo!Answers site a valuable research setting for understanding the general public's online information seeking (Kim & Oh, 2009). The predictors of answers quality in Yahoo!Answers were studied through a comparative analysis (Harper, Raban, Rafaeli, & Konstan, 2008). In another study, topic categories based on answer characteristics and social network interaction were analyzed for 1 month of activities in Yahoo!Answers. This involved 1.2 million questions and 8.5 million answers (Adamic, Zhang, Bakshy, & Ackerman, 2008). The relationship between reward for participants and quality of answers in social Q&A sites like Yahoo!Answers has also been studied. Rewards can make contribution to improvement of answer quality (Gazan, 2011; Shah et al., 2008). To better understand how people seek, share, and evaluate information in a social Q&A environment, Kim, Oh, and Oh (2007) identified the selection criteria people employ when they select best answers in Yahoo!Answers in the context of relevance research. Zhang (2010) explored contextual factors of consumer health information seeking by analyzing health-related questions posted on Yahoo!Answers. The findings show that consumers' questioning behaviors were affected by the knowledge gap, disturbing feelings, and lack of social resources. Questioners tended to have different concerns at different stages of their health and illness. They also had both conceptual and lexical difficulties in formulating questions and sometimes, had high expectations for the quality of the answers. Kim, Oh, and Oh (2008) sampled a total of 700 comments from the Health category in Yahoo!Answers and found utility, socio-emotional value and general statement were the top three criteria in evaluating health-related answers.

Data from social Q&A Websites were used to investigate basic components of image needs of people in daily life (Yoon & Chung, 2011), to examine the dynamics of an online Q&A community based on structuration theory and communities of practice (Rosenbaum & Shachaf, 2010), to evaluate Q&A data quality (Stvilia, Mon, & Yi, 2009), and to study community feedback (Jurczyk & Agichtein, 2007). Another research direction on Q&A is questioners and answerers. Along this line, Gazan (2006) characterizes answerers into specialists and synthesists and found that answers from synthesists were slightly more useful than answers from specialists. Raban (2009) found that the askers present themselves so as to motivate answerers to respond and make an effort to give a good answer while answerers present themselves to maximize the chances for positive feedback. Various research methods have been applied in social Q&A research such as content analysis and correlation analysis (Kim & Oh, 2009), and multinomial logistic tests such as link analysis (Jurczyk & Agichtein, 2007).

Coding analysis has also been applied to Q&A research. This technique can convert a set of data into a meaningful hierarch consisting of a group of subject categories. Hidden themes in the data set can be revealed in the emerging subject structure which can be used to interpret the data set. Coding analysis method has been used for Q&A data analysis to distinguish information questions and conversational questions and assess their archival value (Harper, Moy, & Konstan, 2009). Raban (2009) used the technique to concentrate on the connection between social interaction and the potential value of information in Google Answers. Kelly et al. (2007) used coding as the main categorization strategy to identify evaluation criteria for interactive, analytical Question-Answering systems. Also using the coding method, McCray and Tse (2003) conducted a study based on a data set with 4700 queries from two consumer health sites (ClincialTrials.gov and MEDLINEplus) to analyze search failures, and Keselman, Browne, and Kaufman (2008) utilized the thematic coding method to predict consumer information seeking trajectories in key search strategies.

Information techniques like the MDS (Multiple-Dimensional Scaling) method can be used to effectively illustrate relations among abstract objects, and demonstrate emerging clusters in a data set which is free of any data distributional assumptions. The MDS method can convert a high-dimensional vector space to a low dimensional space where objects in the high dimensional space can be projected onto the low dimensional space. Object relationships can be preserved, and objects and their relationships can be observed and analyzed. Due to the MDS uniqueness, it is applied to many fields of information science such as document co-citation analysis (York, Bohn, Pennock, & Lantrip, 1995), journal co-citation analysis (Hakanen &

Wolfram, 1995), subject co-citation analysis (Small & Garfield, 1985) and Webpage co-citation analysis (Thelwall, 2002; Vaughan, 2006). The MDS method has been used to solve problems in health informatics. Users' search behaviors on obesity were examined, and five related categories (obesity, diet, weight, fat, and food) were discovered in a health portal (Zhang & Wolfram, 2009). In another study, five topic terms (stomach, hip, stroke, depression, and cholesterol) and their related terms were analyzed, and clusters within each of the topic terms were identified in the MDS visual environments respectively (Zhang et al., 2008).

In a study (Poikonen & Vakkari, 2009) a nutrition-related question–answer service designed for diabetics was investigated. The vocabulary differences between lay persons and professionals were examined; term semantic relationships such as equivalence relations, hierarchical relations, associative relations, and causal relations were analyzed; and the overlapping analysis between the term semantic relations from the question–answer data set and two professional thesauri was conducted. In another study (Eerola & Vakkari, 2008) a topic on cardiovascular diseases in a health-related question–answer forum was investigated. In the forum patients posted questions and physicians answered the questions. The differences between semantic expression of consumers and two thesauri on the topic of interest were examined. Zeng et al. (2005) focused their research on consumer-friendly terms identified from a NLM MedlinePlus query log.

In summary, it is quite clear that studies on health consumer term use are many and diverse. They include topics of obesity, cholesterol, nutrition, cardiovascular diseases, etc. The raw data sources for these investigations vary from query transaction logs to a variety of Q&A logs. These research studies have different focuses which include comparisons between consumers' term sematic relations and thesauri, consumer friendly terms, and term clustering analysis. Studies on diabetes term use behaviors based on Q&A data by using the mixed MDS technique and the coding method, however, are scant in the literature.

3. Research method description

Mixed research method (or blended research method, or integrative research method) uses quantitative and qualitative methods, paradigms, and techniques in a complicated study (Hesse-Biber, 2010). In a mixed research method a qualitative research paradigm can be employed in one phase of a research study while quantitative research paradigm can be applied in another phase of the study. Alternately, both qualitative and quantitative research paradigms can be used simultaneously in the same stages of the study. A mixed research method results in a combination of quantitative and qualitative methods which have complementary strengths and which overcomes weaknesses and limitations caused by a single research method. A mixed research method can be used for convergence and corroboration of results from different methods; elaboration, enhancement, illustration, and clarification of results from one method with results from the other method; discovery of contradictions and paradoxes; and expansion of the breadth and width of inquiry (Greene, Caracelli, & Graham, 1989).

In this study, a mixed method of social Q&A log analysis, coding analysis method, and MDS visualization analysis method was used. The first phase used social Q&A log analysis where raw free texts were parsed, terms were extracted, term frequencies were tallied, and associations between terms and records were determined. The second phase employed coding analysis to a group of categories based on the Q&A log analysis results and the third phase applied MDS analysis, a clustering analysis, for the related terms in each of the categories.

3.1. Social Q&A log analysis

Log analysis refers to analysis on data gathered from a server that records users' online activities. The record is known as a log. The data analysis methods which are applied to the log depend on the type of the log and the nature of the log data. Although there is an array of research methods available for users' information seeking behaviors such as survey, interview, observation, experimentation, and query transaction log analysis and each of them has its strengths, the Q&A analysis method in addition to its strength is unique because:

(1) It faithfully records people's real information seeking data. Search tasks are real; information questions are real; and responses are real. In an experimental study the settings and tasks may be designed by researchers.
(2) Q&A data can be extracted and collected within 1 day, 1 week, 1 month, and 1 year from a Q&A log. Robust, reliable, and plausible results should be based on a sufficient data set, and an open social Q&A forum enables researchers to obtain enough data for a study.
(3) Q&A data are detailed. A Q&A record constitutes multiple sentences, paragraphs, and responses, the average number of words and average number of sentences per question in this study were 128.63 and 8.23 respectively. The average number of search terms in a health related query is around 3 according to studies (Spink et al., 2004; White, Dumais, & Teevan, 2008; Zeng, Kogan, Ash, Greenes, & Boxwala, 2002). In a search query, users' information needs have to be generalized and presented by a few keywords. In other words, users' information needs are "compressed" in a relatively short query and as a result, users' information needs may not be fully expressed in a single query. In a Q&A record, users tend to include more information in their information needs question than in a similar keyword search query.

(4) Q&A data may reveal users' motivation and background information which is missing in query data. Although not all Q&A records include detailed users' motivation and background information, a few Q&A records contain this kind of information and other relevant user information. Information like education background, age, gender, reason for asking a question, and users' feelings would be helpful to understanding users' information seeking behaviors. Transaction log analysis based on query terms about users' information seeking analysis lacks this kind of information.
(5) Q&A data have time stamp information. In other words, posting time of a question and responses to the question are recorded and can be traced in a Q&A record. If a research study requires time-sensitive data like a study on flu or allergy, time factor can be a useful dimension for the study.
(6) The Q&A data are diverse and comprehensive. In a Q&A forum, the format, type, and contents of a question from a user are not restricted. Users can post any messages related to a forum theme/topic. As a result, a Q&A log includes almost all aspects of a topic such as diabetes.

The social Q&A log used in this study was Yahoo!Answers (2012). Yahoo!Answers is one of the most influential online social Q&A websites (Alexa, 2012). It includes more than 21 million unique users in the US and 90 million worldwide. It covers 26 categories ranging from Arts & Humanities to Beauty & Style, Social Science to Science & Mathematics, and Health. Each of these categories consists of more detailed subcategories. For instance there are 10 subcategories under Health: Alternative Medicine, Dental, Diet & Fitness, Diseases & Conditions, General Health Care, Men's Health, Mental Health, Optical, Other-Health, and Women's Health. Users can ask and answer questions on any topic in the categories. Advanced search feature enables people to search the Q&A log in a certain period of time, within a certain category, or in a specific field such as a question field, a response field, or both. In this study, the search query used was *Diabetes*, the search was limited to the question field, and length of time was 3 months starting from August 10, 2011 to November 10, 2011. Notice that Yahoo also provides an API to facilitate the Q&A data collection. It can return a structural record where each attribute (question, response, or post time) is identified.

After the data were harvested from the Yahoo Q&A log, a data cleansing process followed. During the data cleansing process, irrelevant and meaningless terms were removed; incomplete, incorrect, and inaccurate terms were modified and fixed; and useful terms were identified and reserved. Using stop lists provided by Fox (1989) the irrelevant and meaningless words were identified for removal. The incomplete, incorrect or inaccurate words refer to misspelled words extracted from the log. These words were identified and corrected by using the editing feature in Microsoft Word. Useful terms were all remaining words.

Each Q&A record received an ID which was used to uniquely identify that record. A record consists of a question and its related responses. All terms from the Q&A records were extracted and tallied. Each term was associated with a corresponding record ID which can trace the term source and term frequencies within a record. All these terms formed a term master file.

Next the term master file was edited for stop words and the remaining terms normalized. Superfluous, useless, and meaningless terms such as "a", "an", "the", "of", and "with" were removed from the term master file because these stop words don't make any contribution to later terms analysis. In the term normalization process, verb forms were reduced to the infinitive, inflected forms of nouns were reduced to the nominative singular, comparatives and superlatives of gradable adjectives were reduced to the absolute form, misspelled terms were corrected, spelling variants were regularized, and apostrophes from the possessive form of words were discarded. The term process normalized the terms to make the later MDS analysis results more reliable and accurate. Finally, terms with low frequencies were also removed from the term master file because these terms made little contribution to the term visualization analyses.

3.2. The data coding method

The data coding method can be used to categorize and classify a set of terms into sub-categories. Sub-categories which emerge as subjects are analyzed in the contexts of raw data, terms are classified, and the topic of a group of related terms is generalized. The data coding analysis leads to a schema which consists of a group of categories. This schema represents the emerging major subject themes and reveals their relationships from the investigated data set. At the beginning of the coding analysis for this study, a simple schema related to diabetes was selected by consulting the website of National Library Medicine (PubMed Health: Diabetes, 2013). The initial schema comprised the following subjects or categories (Causes, Incidence and Risk Factors; Signs and Tests; Treatment; Support Groups; Prognosis; Complications; and Prevention).

Each of the terms in the master file was analyzed based on the contexts of the free-text and then it was classified into a proper category in the schema. As the coding analysis progressed, the schema was constantly revised and adjusted because the existing categories did not fit all the terms in the term master file. As a result, useless categories were removed, new categories were added, and existing categories were modified. After all the terms were analyzed, an updated diabetes schema emerged. The coding analysis process became a process of knowledge discovery.

This coding analysis process prepares the data for an information visualization method such as the MDS method. Visualization methods are usually more suitable with a small or medium size set of data because a large data set would cause an overlapping display of the projected terms within the limited display space. The overlapping display may reduce the effectiveness of the term analysis and even result in a meaningless term analysis. The coding method reduces significantly the number of terms in a category/subcategory if a category/subcategory of the schema rather than the entire schema is selected for visualization analysis. It allows the researchers to conduct an MDS analysis with a reasonable size set of terms

within the category/subcategory at a finer level of granularity and lays a foundation for the effective visualization analysis. The coding analysis for this study was manually conducted by the authors and a medical professional who holds an MD.

3.3. MDS visualization method

Information visualization methods and techniques like the MDS method can effectively project complicated and abstract relationships among objects in a high dimensional space onto a low dimensional space. In the low dimensional space (a two dimensional or three dimensional space), the semantic relationships among the investigated objects can be observed and analyzed, theme patterns among the objects can be formed and discovered, object clusters can be identified and measured, and a rich and holistic picture of the objects can be provided. In addition, some information visualization techniques provide users with an interactive control mechanism. As a result, users can interact with the objects in the visual space, and observe an area of interest from different angles, etc. Due to these unique characteristics, information visualization methods and techniques like the MDS method are particularly suitable for revealing multiple semantic relationships among terms or keywords. The semantic relationships among terms were determined by term association (term co-occurrence) in a Q&A record, term association strength (term co-occurrence frequency), and the ultimate monotonic transformation process of the MDS in which the difference between connections among all the objects (terms) in the high dimensional space and the connections among all the objects (terms) in the low dimensional space reaches a minimum value. The connection between two objects can be measured by one of the *Minkowski* metrics (Zhang, 2008). That is, identified relationships among terms in the low dimensional space primarily depend on the optimal arrangements of the terms defined by a similarity measure in the low dimensional space. However, one weakness of information visualization methods is that the number of the displayed objects in the visual space is limited because the size of a visual space is limited. An overcrowded display in a visual space makes effective visual clustering analysis impossible and even meaningless.

In the MDS analysis process, the dimensionality is reduced, the terms are projected onto a low dimensional space for observation. However, the relationships among the terms in the high dimensional space may not be faithfully illustrated in the low dimensional visual space after the projection. The inconsistence between the relationships among the terms in the high dimensional space and their relationships in the low dimensional space is inevitable. The MDS process minimizes this inconsistence. The degree of the inconsistence can be effectively measured by the stress value (S), which is defined in Eq. (1). The smaller a stress S is, the better the MDS result is; and vice versa. The stress value is used to measure the quality of an MDS result.

$$S = \left(\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (f(T_i, T_j) - D(T_i, T_j)^2}{\sum_{i=1}^{n} \sum_{j=1}^{n} (D(T_i, T_j)} \right)^{1/2} \tag{1}$$

In Eq. (1) n is equal to the number of all terms involved, $D(T_i,T_j)$ is the Euclidean distance between two terms T_i and T_j in the low visual space, $f(T_i,T_j)$ is the similarity between terms T_i and T_j in the high dimensional space, i and j are two indexes.

The MDS requires a term–term proximity matrix as an input. The proximity matrix is an $n * n$ matrix that describes the relationships among terms. The MDS analysis result provides a visual display where the projected terms are observed, terms are clustered, and term relationships are discovered.

The MDS analysis can be enhanced using a hierarch clustering method which helps researchers identify and confirm resultant clusters from the MDS analysis. In the MDS visual display, some terms with similar characteristics may be projected onto the same spot in the MDS display. As a result, some of the overlapping terms may not be shown in the display. It causes unnecessary difficulty for clustering analysis in the MDS display. With the help of a hierarch clustering method, this problem can be easily solved. Secondly, in the MDS display all terms are presented in a 3-D space and the boundary between two clusters may not be clear if a viewing angle is not properly selected. A hierarch clustering method assists the researchers in identifying the boundary between the clusters and confirms the results from the MDS analysis. In this study the hierarch clustering method was applied to the same data set using the between-groups linkage clustering method.

The MDS visualization analysis in this study consisted of two levels: visualization analysis with each of categories at a term level which enabled users to observe semantic relationships among terms within a category; and visualization analysis for all the categories at a category level. It demonstrated overall relationships among the identified categories. The software used for an MDS analysis was SPSS (Version 20).

3.3.1. Term visualization analysis within each of the categories

After the coding process, a group of categories emerges (see Eq. (2)). Here r is the number of the categories.

$$Schema = \{C1, C2, C3, \ldots, Cr\} \tag{2}$$

Each category contains a set of related terms which can form a term–record matrix where its columns are Q&A records while its rows are terms. For instance, Cq is a category ($1 \leqslant q \leqslant r$), and it can be represented in Eq. (3). In Eq. (3) n is the number of the Q&A records and k is the number of the terms in category Cq. Here a_{ij} is a cell in the matrix and the value of a_{ij} indicates frequency of the term i in the Q&A record j.

$$M_{Cq} = \begin{pmatrix} a_{11} & .. & .. & a_{1n} \\ a_{21} & .. & .. & a_{2n} \\ & \cdots & \cdots & \\ a_{k1} & .. & .. & a_{kn} \end{pmatrix} \tag{3}$$

In order to achieve a sound MDS result in this study, M_{cq} was simplified by removing some columns with low frequencies. If the sum of cells in a column was lower than a predetermined frequency cutoff point, the column was removed from the matrix.

This term–record matrix was converted to a term–term proximity matrix based on a similarity measure. There are many similarity measures available for this purpose. Selection of a similarity measure has a strong impact on the stress value of an MDS analysis result. The distance-based similarity measure (Zhang & Rasmussen, 2001) and Cosine similarity measure were considered in this study because of their good performances in a pilot study.

$$PM_{Cq} = \begin{pmatrix} b_{11} & .. & .. & b_{1k} \\ b_{21} & .. & .. & b_{2k} \\ & \cdots & \cdots & \\ b_{k1} & .. & .. & b_{kk} \end{pmatrix} \tag{4}$$

The proximity matrix PM_{cq} is a $k * k$ symmetric matrix and it served as input data for the MDS analysis.

3.3.2. Category visualization analysis among all categories

The category visualization analysis among all the categories provides an overall and holistic picture for all the emerged categories. The term visualization analysis within each of categories was conducted at a term level while the category visualization analysis among all categories was conducted at a category level. It was apparent that in this study the objects displayed and analyzed in the MDS configuration were the emerging categories rather than individual terms. To visualize these categories, the surrogates of these categories had to be clearly defined and properly represented in the high dimensional space.

A centroid of a cluster in a high dimensional space can be defined to represent that cluster (Zhang & Wolfram, 2001). A centroid is located at the "center" of the cluster in the high dimensional space. If an identified category in the coding analysis is treated as a cluster in the high dimensional vector space, its centroid can be employed to represent it in the space.

For the category *Cq*, its centroid can be defined as follows:

$$Centroid\_M_{Cq} = \left(\frac{\sum_{i=1}^{k} a_{i1}}{k}, \quad \frac{\sum_{i=1}^{k} a_{i2}}{k}, \quad \ldots, \quad \frac{\sum_{i=1}^{k} a_{in}}{k} \right) \tag{5}$$

Similarly, all the categories can be represented by their centroids using Eq. (5).

Using the centroid for each category, a category–record matrix was created and used to generate a category–category proximity matrix.

$$PM_{Category} = \begin{pmatrix} b_{11} & .. & .. & b_{1r} \\ b_{21} & .. & .. & b_{2r} \\ & \cdots & \cdots & \\ b_{r1} & .. & .. & b_{rr} \end{pmatrix} \tag{6}$$

Here r is the number of the categories. An $r * r$ symmetric matrix is formed with each row in the matrix representing a category. This matrix served as the input of the later MDS analysis for category visualization among all categories.

4. Results and discussions

4.1. Analysis of the collected data

The social Q&A log in Yahoo!Answers was used in this study. The formulated query was *Diabetes*. Thanks to the advanced search features in Yahoo!Answers the search was limited to the question field within the Health category. Data was collected between August 10, 2011 and November 10, 2011. As a result, there were 2604 retrieved records returned from Yahoo!Answers. Records completely not related to diabetes were deleted leaving 2565 records. Each record consisted of a question and its related responses. Usually the question included a question title and explanation. The total number of extracted words from the records was 1,043,158. This word set contained all the words extracted from the log, including duplicate words. The descriptive statistics of the returned record set are summarized in Tables 1 and 2. In Tables 1 and 2, the words in a question or a response include both keywords and stop-words.

According to the statistical data, there were 150 records which only included question titles without detailed explanations. They accounted for 5.85% of the investigated records. The longest sentence in questions had 1055 words while the shortest sentence only had 3 words.

Table 1
The descriptive statistics of questions in the returned record set.

Description	Result
Average number of words	128.63
Median	97
Minimum number of words	3
Maximum number of words	1055
Standard deviation of words	120.25
Average number of sentences	8.23
Median	6
Minimum number of sentences	1
Maximum number of sentences	73
Standard deviation of sentences	7.36

Table 2
The descriptive statistics of responses in the returned record set (Part I).

Description	Result
Average number of words	254.83
Median	168.5
Minimum number of words	0
Maximum number of words	2117
Standard deviation of words	267.42
Average number of sentences	16.01
Median	11
Minimum number of sentences	0
Maximum number of sentences	119
Standard deviation of sentences	15.99

Yahoo Q&A allows questioners to add more information to original questions if the original questions are not clearly defined in the first place. According to this study's statistical data, there were 23 questions which contain additional information added by questioners. They account for 0.009% of the total questions.

In this study some responses to questions contained information resources. According to the statistical data, there were 369 responses which contain information resources which accounted for 14.39% of the total questions. These resources include http addresses of authoritative information such as (www.diabetes.org, http://www.youtube.com/watch?-v=G3rvItKih...), and relevant books such as *New Choices in Natural Healing Prevention Magazine Health Books* and *Harrison Book of Internal Medicine*.

The average number, median, minimum number, maximum number, and standard deviation of responses in a record are shown in Table 3.

Two *t*-tests were conducted to ascertain the differences between the number of words in questions and the number of words in responses ($T1$), and the differences between the number of sentences in questions and the number of sentences in responses ($T2$) respectively. Significance level for both the tests was 0.05. The resultant *p*-values for both $T1$ and $T2$ are 0.000 in Table 4. It suggests that there are significant differences between the number of words in questions and the number

Table 3
The descriptive statistics of responses in the returned record set (Part II).

Description	Result
Average number of responses	2.99
Median	3
Minimum number of responses	0
Maximum number of responses	24
Standard deviation of responses	2.02

Table 4
Results of the two *t*-tests.

T-test	df	*P* value
The number of words in questions vs. the number of words in responses (T1)	2564	0.000
The number of sentences in questions vs. the number of sentences in responses (T2)	2564	0.000

of words in responses at the significance level 0.05, and there are significant differences between the number of sentences in questions and the number of sentences in responses at the significance level 0.05.

A total of 20,000 unique words were extracted from the returned Q&A records in Yahoo!Answers. The number of the stop-words was 652. The cutoff point for low frequency terms was set to 3 resulting in the removal of 12,643 low frequency terms. After discarding the stop-words and terms with low frequencies, the remaining 7357 words were normalized to 3979 keywords. This list was further reduced with a keyword cutoff point set to 5, making the number of keywords in each category manageable for visualization analysis.

The term master file finalized at 1106 words and its descriptive statistics are shown in Table 5. The top 20 terms with the highest frequencies are listed in Table 6. It is no surprise that the term diabetes is located at the top of the list. Four terms (eat, food, drink, diet) are related to *Nutrient*, and 9 terms (blood, sugar, test, high, body, glucose, level, low) are related to *Test*.

Term growth analysis allows people to observe a relationship between the number of unique words extracted from the Q&A records and the number of the processed Q&A records. The analysis helps the researchers determine the length of the time period within which the Q&A records are retrieved. In Fig. 1, the *X*-axis is the number of the Q&A records and the *Y*-axis is the number of unique terms. The curve in the figure illustrates that when the number of the Q&A records reaches 2,500, the number of the unique words in the master term file has stabilized. This analysis confirms that the 2 month time period of data collection in this study was sufficient to obtain enough words for the term analysis.

Table 5
The descriptive statistics of term frequencies in the term master file.

Categories	Result
Average term frequency	179.23
Median	42
Minimum term frequency	6
Maximum term frequency	10,716
Standard deviation of term frequency	555.91

Table 6
Top 20 most frequent terms of the term master file.

Terms	Frequency	Terms	Frequency
Diabetes	10,716	Cause	2341
Blood	6267	Symptom	2328
Sugar	6009	Diabetic	2265
Eat	4591	Body	2222
Doctor	3955	Drink	2098
Test	3541	Diet	2054
Weight	2624	Food	1990
Feeling	2508	Glucose	1986
Insulin	2498	Level	1958
High	2488	Low	1922

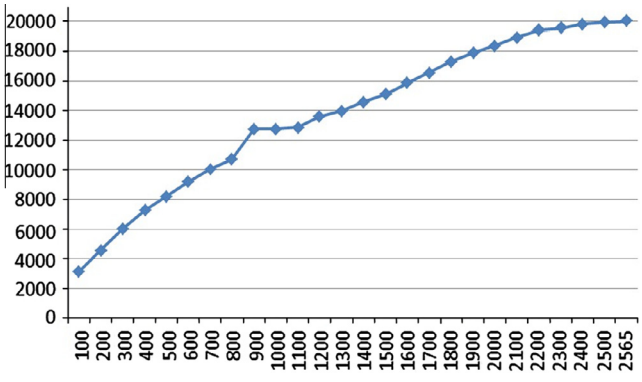

Fig. 1. The term growth chart.

4.2. Coding analysis results

After removing useless words and words with low frequencies, normalizing (see previous section) the term master file underwent further data coding analysis. Each individual term in the term master file was examined, analyzed, and mapped onto a proper category in the predefined schema of seven categories (*Causes*, *Incidence*, *and Risk Factors*; *Signs and Tests; Treatment; Support Groups*; *Prognosis*; *Complications*; and *Prevention*). The schema was revised, and adjusted during the data coding analysis process, and a new schema for diabetes, based on users' Q&A behaviors, emerged (see Fig. 2). The newly emerged schema has 12 categories. It is not surprising that in comparison with the original schema, user-oriented categories like *Social and Culture*, *Nutrient*, *Affect*, and *Life Style* were added to the new schema. Note that the former category *Support Groups* was revised to a more specific *Education & Info Resource*, and the former category *Prevention* was discarded.

The sizes and percentages of the 12 categories are shown in Table 7 and the display of the percentages of the 12 categories is demonstrated in Fig. 3. The category *Nutrient* clenches the first position and accounts for 17.31%; *Diagnosis & Test* claims the second position (for 11.39%); and *Complication & Related Disease* holds the third position (for 8.88%). These dominant categories reflect primary concerns about diabetes by the common health consumers. The category *Nutrient* at the top of the ranked list suggests that food related questions are popular and they have a very close relationship to diabetes.

In Table 7, size refers to the total number of unique terms in that category and percentage refers to the per cent of the total number of unique terms in that category to the total number of unique terms of the master file.

4.3. The MDS visualization analysis results

The visualization analyses on diabetes related terms from the Q&A log were conducted at two levels based on the emerging diabetes schema: term analysis within each of the categories or subcategories; and category analysis within the schema. These two levels of visualization analysis are discussed separately.

4.3.1. Within a category

There are 12 primary categories within the schema in Table 7. Notice that category 3 (*Diagnosis and Test*) and category 12 (*Nutrient*) are two large categories in the schema. Category 3 and category 12 contain 127 and 193 terms respectively. These two large categories were treated differently in the term analysis within each of the categories. Category 3 was broken down into two subcategories (3.1 *Diagnosis* and 3.2 *Test*) and category 12 was divided into four subcategories (12.1 *Vegetable*, *Fruits & Grain*, 12.2 *Meat*, *Seafood & Dairy*, 12.3 *Fast Food*, *Drink & Condiment*, and 12.4 *Miscellaneous*) to avoid overwhelming term displays in the visual spaces.

4.3.1.1. Cause & Pathophysiology. In the category *Cause & Pathophysiology*, the Cosine similarity measure was used to create the term–term proximity matrix. In the MDS analysis, the *Minkowski* distance measure was employed and the *Minkowski*

Diabetes		
	1. Cause & Pathophysiology	
	2.Sign & Symptom	
	3.Diagnosis & Test	3.1 Diagnosis
		3.2 Test
	4. Organ & Body Part	
	5. Complication & Related Disease	
	6. Medication	
	7. Treatment	
	8. Education & Info Resource	
	9. Affect	9.1 Emotion
		9.2 Feeling
	10. Social & Culture	10.1 Family
		10.2 Social
		10.3 Religion
	11. Lifestyle	
	12. Nutrient	12.1 Vegetable, Fruit & Grain
		12.2 Meat, Seafood & Dairy
		12.3 Fast Food, Drink & Condiment
		12.4 Miscellaneous

Fig. 2. The schema for diabetes.

Table 7
The sizes and percentages of the 12 categories and subcategories.

Categories	Size	Percentage	Number of Records
1. Cause & Pathophysiology	73	6.60	1806
2. Sign & Symptom	95	8.59	1779
3. Diagnosis & Test	127	11.48	
3.1. Diagnosis	71	6.42	2225
3.2. Test	56	5.06	1660
4. Organ & Body Part	66	5.97	1768
5. Complication & Related Disease	91	8.23	730
6. Medication	74	6.69	548
7. Treatment	71	6.42	1403
8. Education & Info Resource	88	7.96	970
9. Affect	77	6.96	1237
10. Social & Culture	83	7.50	1157
11. Lifestyle	68	6.15	927
12. Nutrient	193	17.45	
12.1. Vegetable, Fruits & Grain	76	6.87	407
12.2. Meat, Seafood & Dairy	27	2.44	294
12.3. Fast Food, Drink & Condiment	41	3.71	836
12.4. Miscellaneous	49	4.43	1568
Total	1106	100.00	

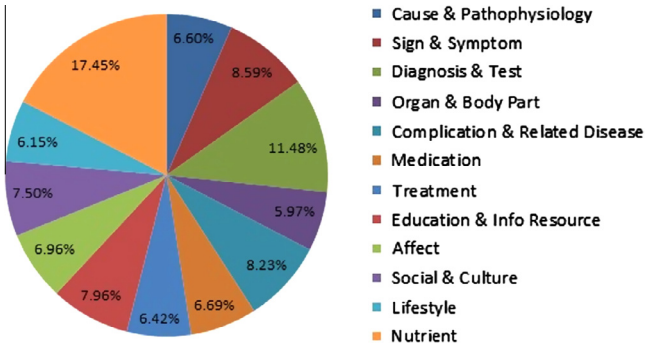

Fig. 3. The display of the percentages of the 12 categories.

power was equal to 1. The resultant stress value was 0.10238 and the corresponding *RSQ* was equal to 0.94986. Three clusters were identified in the MDS visual space (Fig. 4). There were 7, 20, and 46 terms in the three clusters respectively. The detailed terms of the three clusters are listed in Table 8.

It is evident that Cluster 1 identifies the main factors which cause diabetes. They are overweight, obesity, gene, and age. Cluster 2 shows the pathophysiology of type I diabetes mellitus characterized by loss of the insulin-producing beta cells of the islets of Langerhans in the pancreas resulting in insulin shortage, and type II diabetes mellitus characterized by insulin resistance. Cluster 3 provides other information on the cause and pathophysiology of diabetes such as stress, environment, pressure, heredity, and immune deficiency.

4.3.1.2. Sign & Symptom. In the category Sign & Symptom, the Cosine similarity measure was used to create the term–term proximitymatrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS visual space (Fig. 5). There were 7, 10, 49 and 29 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 9.

Cluster 1 primarily reflects symptoms related to cold and cough. Cluster 2 includes two groups: one is related to sleep (night, sleep, and wake) and the other is related to physical symptoms (twitch, tremor, shake, and faint).

4.3.1.3. Diagnosis. In the category *Diagnosis*, the Cosine similarity measure was used to generate the input matrix. The Minkowski power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS visual space (Fig. 6). There were 6, 7, 39 and 19 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 10.

In Cluster 1 cellular metabolism falling is one of diagnostic results of thirst which is a common symptom of diabetes. In Cluster 2, LADA (Latent Autoimmune Diabetes of Adults) and MODY (Maturity Onset Diabetes of the Young) are identified as different types of diabetes mellitus. DI (Diabetes Insipidus) is also a kind of diabetes. Two special types of diabetes GDM

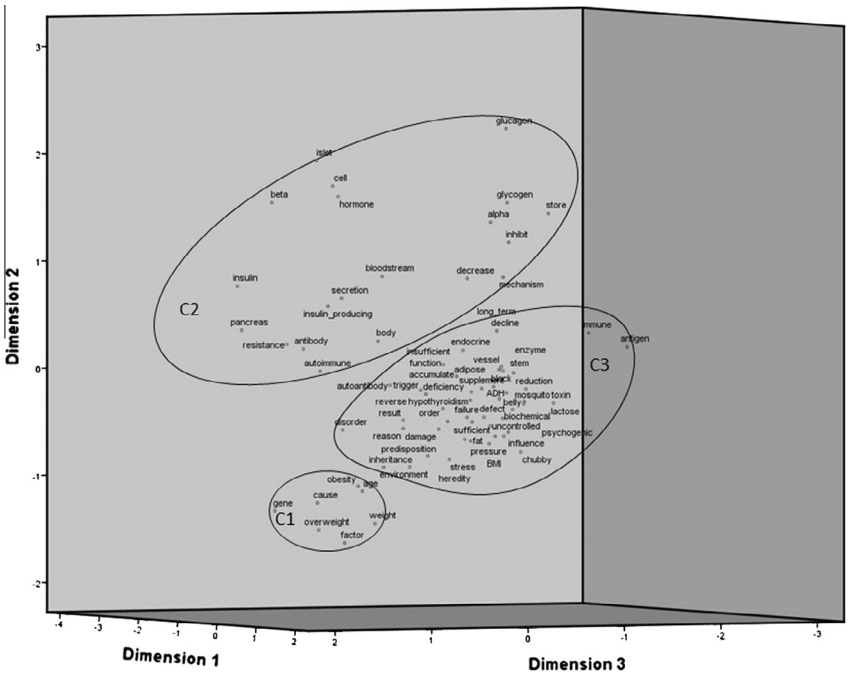

Fig. 4. The MDS display of Cause & Pathophysiology.

Table 8
Result of Cause & Pathophysiology.

Cluster	Result
C1	Overweight, obesity, factor, gene, age, weight, cause
C2	Alpha, glucagon, glycogen, store, inhibit, decrease, mechanism, body, cell, hormone, beta, islet, secretion, insulin, resistance, pancreas, insulin-producing, bloodstream, antibody, autoimmune
C3	Endocrine, disorder, deficiency, sufficient, supplement, reverse, pressure, uncontrolled, stress, block, order, result, damage, vessel, function, failure, stem, decline, heredity, inheritance, environment, predisposition, lactose, trigger, chubby, influence, accumulate, BMI, biochemical, insufficient, long-term, psychogenic, belly, ADH, enzyme, mosquito, defect, adipose, antigen, toxin, autoantibody, hypothyroidism, reduction, immune, fat, reason

Note: ADH (antidiuretic hormone); BMI (Body Mass Index).

(Gestational Diabetes Mellitus) and IDDM (Insulin Dependent Diabetes Mellitus) were also identified in Cluster 3. In Cluster 4, pre-diabetes, non-diabetes, diabetes type I, and diabetes type II were found.

4.3.1.4. Test. In the category *Test*, the Cosine similarity measure was used to produce input data. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS space (Fig. 7). There were 6, 5, 12 and 33 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 11.

Cluster 1 contains basic diabetes test indicators. In Cluster 2 WBC test can be used to measure monocytes or white blood cells. In Cluster 3, A1c (HbA1c), GTT (Glucose Tolerance Test), OGTT (Oral Glucose Tolerance Test), and TSH (thyroid-stimulating hormone) are common diabetes tests. In Cluster 4, EKG (Elektrokardiogramm), EGFR (Estimated Glomerular Filtration Rate), GFR (Glomerular Filtration Rate), BMR (Basal Metabolic Rate), and CBC (Complete Blood Count) are also tests related to diabetes.

4.3.1.5. Organ & Body Part. In the category *Organ & Body Part*, the distance similarity measure ($c = 1.3$, and $k = 2$) was used to generate input data. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Three clusters were identified in the MDS space (Fig. 8). There were 41, 10 and 15 terms in the three clusters respectively. The detailed terms of the three clusters are listed in Table 12.

In Cluster 1, some body parts are clearly associated with diabetic symptoms such as dry mouth, eye floaters, and blurry eye. The terms leg, thigh, knee, hand, shoulder, arm, muscle, neck, wrist are grouped in Cluster 2. With diabetes these body parts frequently suffer from a burning sensation, tingle, and numbness. Cluster 3 includes some of the body parts that are related to complications (for instance, heart attacks and kidney failure).

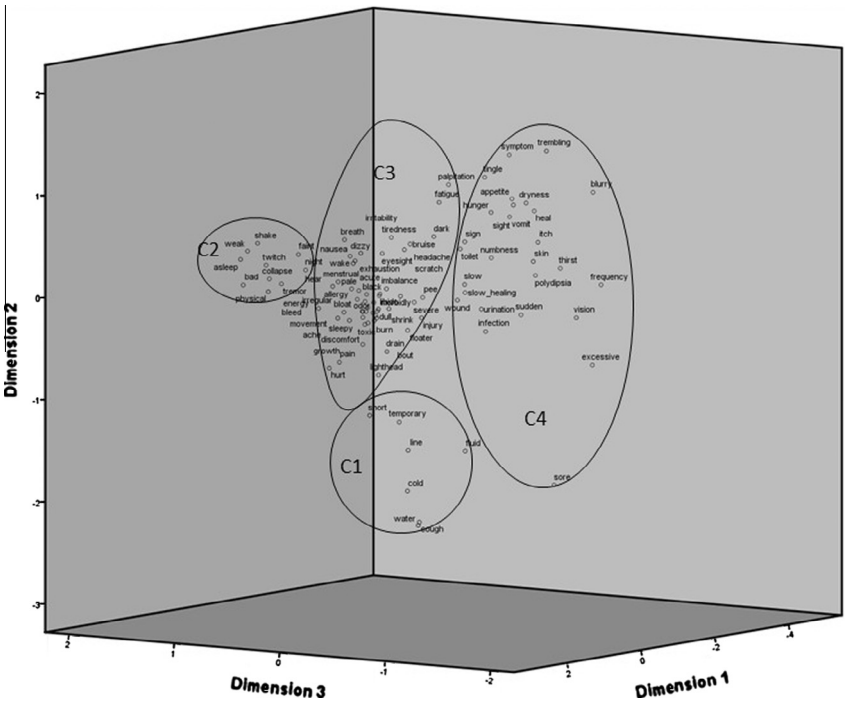

Fig. 5. The MDS display of Sign & Symptom.

Table 9
Result of Sign & Symptom.

Cluster	Result
C1	Cold, water, cough, fluid, line, short, temporary
C2	Asleep, wake, night, physical, weak, collapse, twitch, tremor, shake, faint
C3	Discomfort, bout, morbidly, skinny, odor, drain, irritability, exhaustion, pale, allergy, dull, burn, energy, toxic, growth, acute, imbalance, bleed, movement, swollen, ache, sleepy, thin, hair, menstrual, irregular, bloat, eyesight, black, scratch, lighthead, shrink, libido, star, floater, injury, severe, hurt, pain, hear, bad, toilet, dark, pee, fatigue, tiredness, nausea, headache, dizzy, breath
C4	Tingle, numbness, heal, slow, dryness, skin, itch, infection, blurry, vision, sudden, sign, frequency, thirst, symptom, excessive, urination, hunger, polydipsia, sore, appetite, vomit, trembling, sight, palpitation, bruise, slow-healing, wound

4.3.1.6. Complication & Related Disease. In the category *Complication & Related Disease*, the distance similarity measure ($c = 1.3$, and $k = 2$) was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS space (Fig. 9). There were 30, 24, 12 and 25 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 13.

In Cluster 1, cataracts, retinopathy, bulimia, and colitis are complications. Cataracts and retinopathy are eye-related complications. Schizophrenia, disability, hypochondria, Alzheimer, bipolar, trauma, and migraine are diabetes-related diseases. Schizophrenia, hypochondria, bipolar, and trauma are also related to mental diseases. In Cluster 2, ketoacidosis, edema, glaucoma, neuropathy, and polyuria are complications, while COPD (chronic obstructive pulmonary disease), anorexia, fibromyalgia, and Lyme disease are diabetes-related diseases. In Cluster 3, DKA is a complication. In Cluster 4, skin-related terms are identified: anemia, acanthosis, nigricans, swell, gangrene, eczema, and ulcer. The complications are proteinuria, polyphagia, acanthosis, nigricans, and blindness.

4.3.1.7. Medication. In the category *Medication*, the distance similarity measure ($c = 1.3$, and $k = 2$) was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Five clusters were identified in the MDS space (Fig. 10). There were 31, 16, 10, 9 and 8 terms in the five clusters respectively. The detailed terms of the five clusters are listed in Table 14.

Cluster 1 includes medication for stress and depression (fluoxetine, effexor, magnesia), for suppression of urine (antidiuretic, gabapentin), inflammation (amoxicillin), and for diabetes type II (vasopressin, glucovance, glucophage, onglyza, avandia). Cluster 2 includes medication for psychosis and schizophrenia (antipsychotic, Seroquel, Lexapro), for pain relief (aspirin, ibuprofen), and for diabetes type I (Novolog, Humalog). Cluster 3 contains additional medication for the diabetes type II

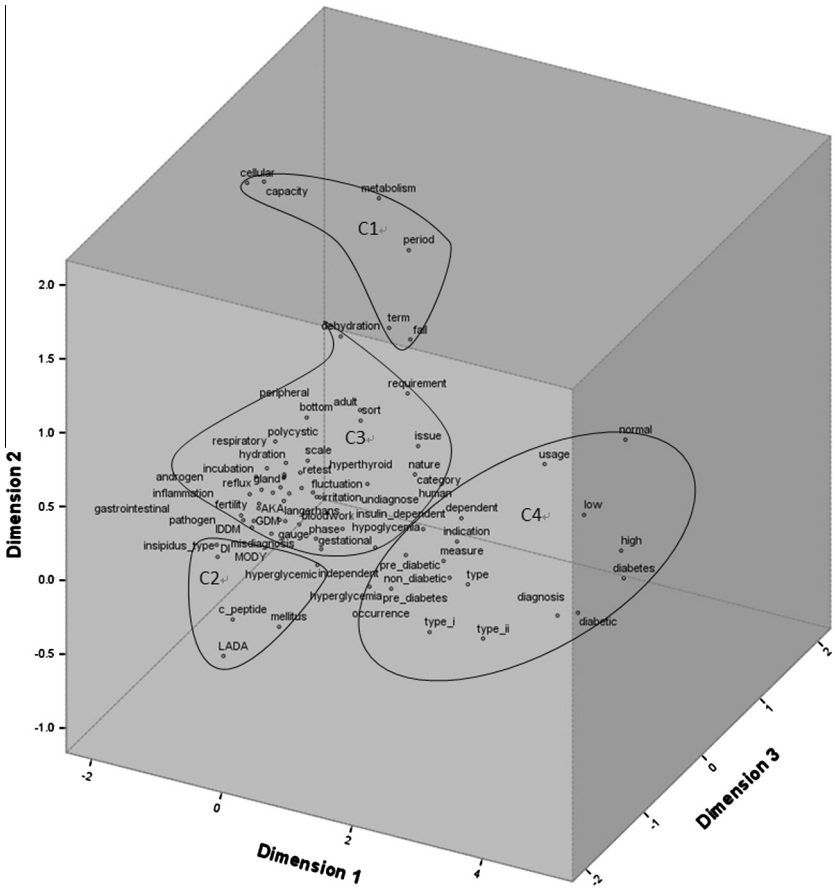

Fig. 6. The MDS display of Diagnosis.

Table 10
Result of Diagnosis.

Cluster	Result
C1	Capacity, cellular, metabolism, period, fall, term
C2	c-Peptide, LADA, mellitus, insipidus-type, DI, MODY, independent
C3	GDM, gestational, IDDM, insulin-dependent, androgen, polycystic, gauge, phase, misdiagnosis, AKA, hyperglycemic, human, incubation, issue, nature, inflammation, reflux, sort, dehydration, hydration, bottom, scale, undiagnosed, vein, peripheral, gland, hyperthyroid, irritation, retest, bloodwork, fluctuation, category, Langerhans, fertility, gastrointestinal, respiratory, pathogen, adult, requirement
C4	Type-I, type-II, diabetes, type, high, diabetic, low, diagnosis, normal, usage, dependent, occurrence, hypoglycemia, hyperglycemia, pre-diabetes, pre-diabetic, non-diabetic, indication, measure

(actos, acarbose). It is no surprise because Cluster 2 is adjacent to Cluster 3 in the visual space. Cluster 5 involves medication for both the diabetes type I (Lantus) and the diabetes type II (metformin).

4.3.1.8. Treatment. In the category *Treatment*, the Cosine similarity measure was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS space (Fig. 11). There were 6, 55, 4 and 6 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 15.

Cluster 1 includes people related to diabetes such as caregivers, therapist, psychologist, psychiatrist, and MD. Cluster 2 includes another group of medical practitioner related to diabetes (doctor, nurse, neurologist, dentist, specialist, cardiologist, endocrinologist, and dermatologist), medical treatment places (clinic, emergency, hospital, ER, ICU, pharmacy, pediatric hospital), and medical equipment (ambulance, pacemaker, monitor). Cluster 3 includes surgery and surgeons. Complementary therapy and other means for treating diabetes are shown in Cluster 4.

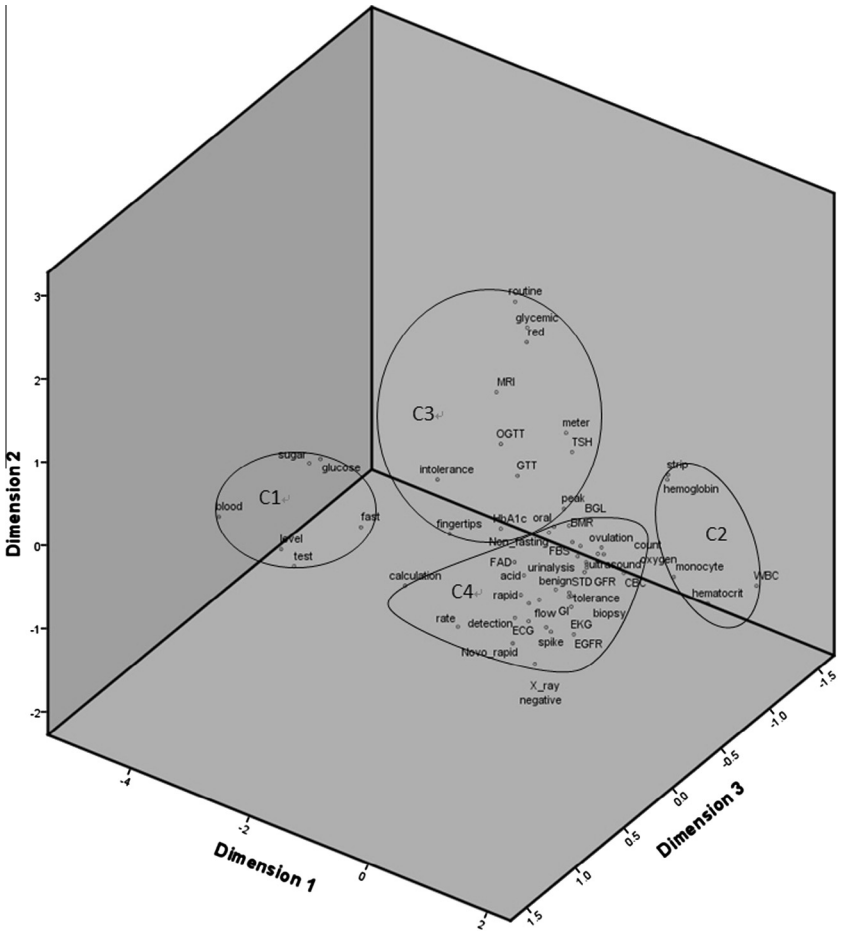

Fig. 7. The MDS display of Test.

Table 11
Result of Test.

Cluster	Result
C1	Sugar, blood, level, glucose, fast, test
C2	Monocyte, strip, hematocrit, WBC, hemoglobin
C3	Peak, MRI, meter, TSH, HbA1c, GTT, OGTT, intolerance, fingertips, glycemic, red, routine
C4	Flow, detection, calculation, CBC, BP, EGFR, GFR, urinalysis, biopsy, tolerance, Novo-rapid, rapid, spike, benign, ECG, BMR, rate, acid, Non-fasting, FAD, negative, BGL, minimum, ovulation, oral, GI, ultrasound, oxygen, count, X-ray, EKG, FBS, STD

Note: STD (STD test), WBC (white blood cells test), GI (Glycemic Index), MRI (Magnetic Resonance Imaging), FAD (Flavin Adenine Dinucleotide).

4.3.1.9. Education & Info Resource. In the category *Education & Info Resource*, the distance similarity measure ($c = 1.3$, $k = 2$) was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Five clusters were identified in the MDS space (Fig. 12). There were 6, 21, 24, 15, and 22 terms in the five clusters respectively. The detailed terms of the five clusters are listed in Table 16.

Two major national medical institutes, the ADA (American Diabetes Association) and the FDA (Food and Drug Administration) are shown in Cluster 1. Cluster 2 includes newly emerging media (blog, email, YouTube, Wikipedia) and traditional media (disk, TV, video). Cluster 3 includes educational places (universities, colleges, and education centers), online information search channels (search, Yahoo, Google), and educational sources (reading, news, movie saying, tip, course). Cluster 4 includes education (school, class, study, consultation) and Internet (post, website, link, index). New information technology (Facebook, twitter, app, iPhone), and traditional communication (letter, on chat, mail, speech, press) appear in Cluster 5.

4.3.1.10. Affect. In the category *Affect*, the distance similarity measure ($c = 1.3$, and $k = 2$) was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the

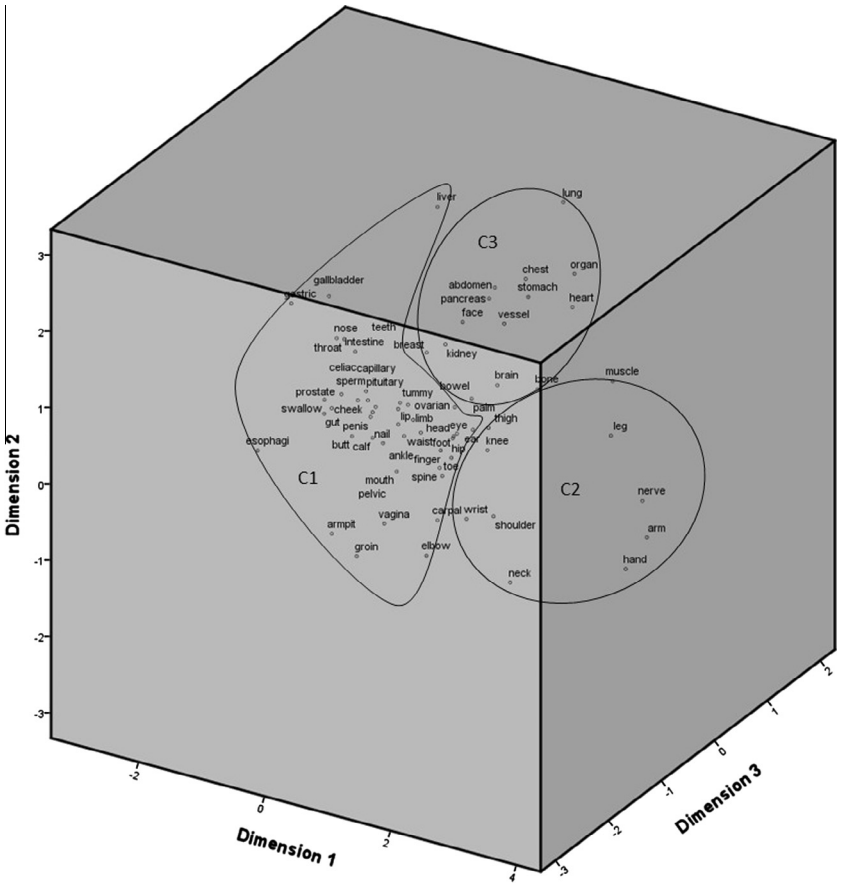

Fig. 8. The MDS display of Organ & Body Part.

Table 12
Result of Organ & Body Part.

Cluster	Result
C1	Throat, nose, vagina, groin, mouth, armpit, liver, gallbladder, esophagi, swallow, gastric, carpal, elbow, ankle, calf, hip, spine, waist, ovarian, pituitary, limb, eye, foot, toe, nail, prostate, pelvic, tummy, teeth, intestine, celiac, head, ear, cheek, butt, gut, penis, sperm, finger, lip, capillary
C2	Nerve, leg, thigh, knee, hand, shoulder, arm, muscle, neck, wrist
C3	Lung, chest, heart, breast, brain, bone, vessel, organ, palm, face, pancreas, stomach, abdomen, bowel, kidney

MDS space (Fig. 13). There were 32, 9, 12 and 24 terms in the four clusters respectively. The detailed terms of the five clusters are listed in Table 17.

This category covers a lot of terms about positive and negative emotions, feelings, and attitudes caused by diabetes in all 4 clusters. It is clear that the negative terms outnumber the positive terms. There are only a few positive terms (excited, easy, love, cheer, happy, cool, enjoy). Most of these positive terms are shown in Cluster 4 with a few also in Clusters 1 and 3.

4.3.1.11. Social & Culture. In the category *Social & Culture*, the Cosine similarity measure was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Six clusters were identified in the MDS space (Fig. 14). There were 9, 4, 5, 39, 15 and 11 terms in the six clusters respectively. The detailed terms of the six clusters are listed in Table 18.

Clusters 1 and 2 terms relate to religion. It is interesting that it is hard to find other religions except Christianity (Jesus, Christian, Christ, Bible) in this category. Cluster 3 basically addresses marriage (divorce, husband, marriage, housekeeper, adoption). Cluster 4, the largest cluster in this category, includes family (relative, grandfather, senior, generation, elderly, grandmother, aunt) and profession (professor, teacher, actress, military). Clusters 5 and 6 which are adjacent to each other in the visual space contain terms related to family and relatives.

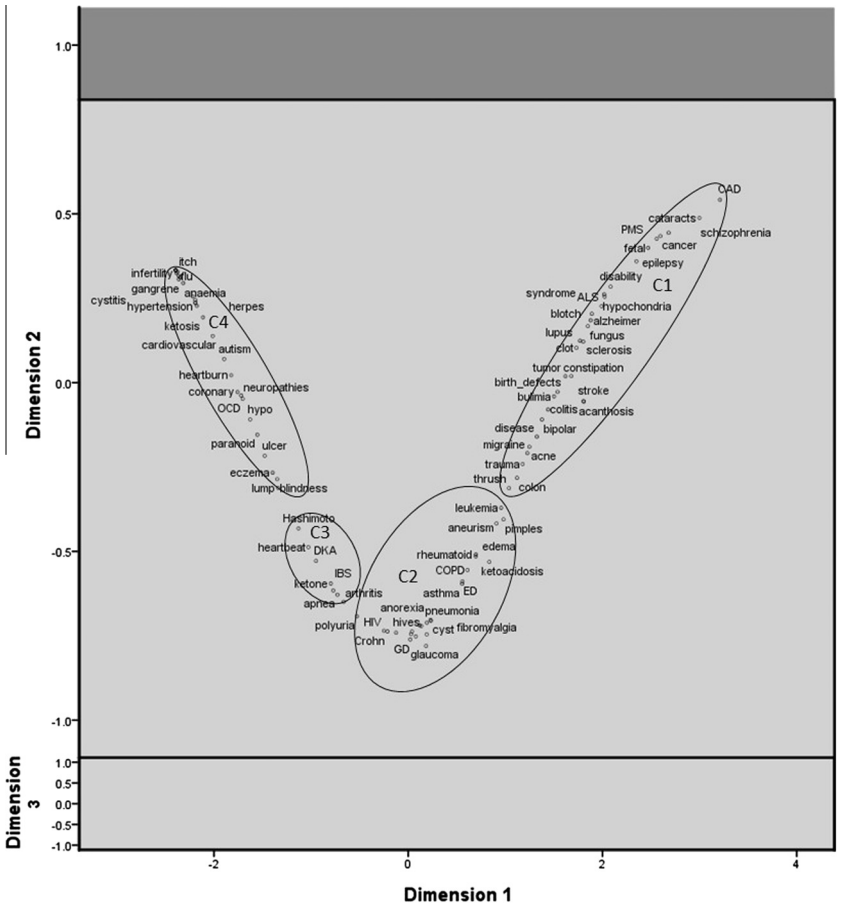

Fig. 9. The MDS display of Complication & Related Disease.

Table 13
Result of Complication & Related Disease.

Cluster	Result
C1	CAD, cataracts, schizophrenia, PMS, cancer, fetal, epilepsy, disability, hypochondria, ALS, Alzheimer, blotch, fungus, lupus, constipation, clot, sclerosis, tumor, syndrome, birth-defects, bulimia, colitis, bipolar, disease, trauma, thrush, colon, acne, migraine, retinopathy, stroke
C2	Ketoacidosis, ED, asthma, COPD, edema, leukemia, rheumatoid, cyst, anorexia, pneumonia, fibromyalgia, GD, scab, glaucoma, hives, Crohn, HIV, aneurism, pimples, mental, PCOS, neuropathy, Lyme, polyuria
C3	Ketone, arthritis, IBS, apnea, DKA, heartbeat, Hashimoto
C4	Hypo, paranoid, heartburn, OCD, coronary, neuropathies, cardiovascular, anemia, herpes, cystitis, hypertension, gangrene, flu, chronic, autism, ketosis, proteinuria, polyphagia, seizure, complication, infertility, acanthosis, nigricans, swell, sinus, eczema, ulcer, lump, blindness

Note: ALS (Amyotrophic Lateral Sclerosis, CAD (Coronary Artery Disease), GD (Gaucher Disease), IBS (Irritable Bowel Syndrome), PCOS (Polycystic Ovarian Syndrome), PMS (Premenstrual Syndrome), Hashimoto (Hashimoto's Autoimmune Thyroiditis), ED(Erectile Dysfunction), DKA (Diabetic Ketoacidosis).

4.3.1.12. Lifestyle. In the category *Lifestyle*, the Cosine similarity measure was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS space (Fig. 15). There were 16, 16, 32 and 4 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 19.

The terms in Cluster 1 are general terms on lifestyle and also include terms related to walking (trip, treadmill, jog). The terms in Cluster 2 are related to sports (basketball, tennis, play, sport, football, soccer, bike, athletic). Cluster 3, the largest one in this category covers a variety of topics such as sports (Olympic, marathon, racing, runner, climb, swim), entertainment (park, poke, ballet, band, rap, music), fitness (bodybuilding, regimen), and sleep (bed, sleep, bedtime, nap). Cluster 4 uniquely includes Tai-Chi and yoga.

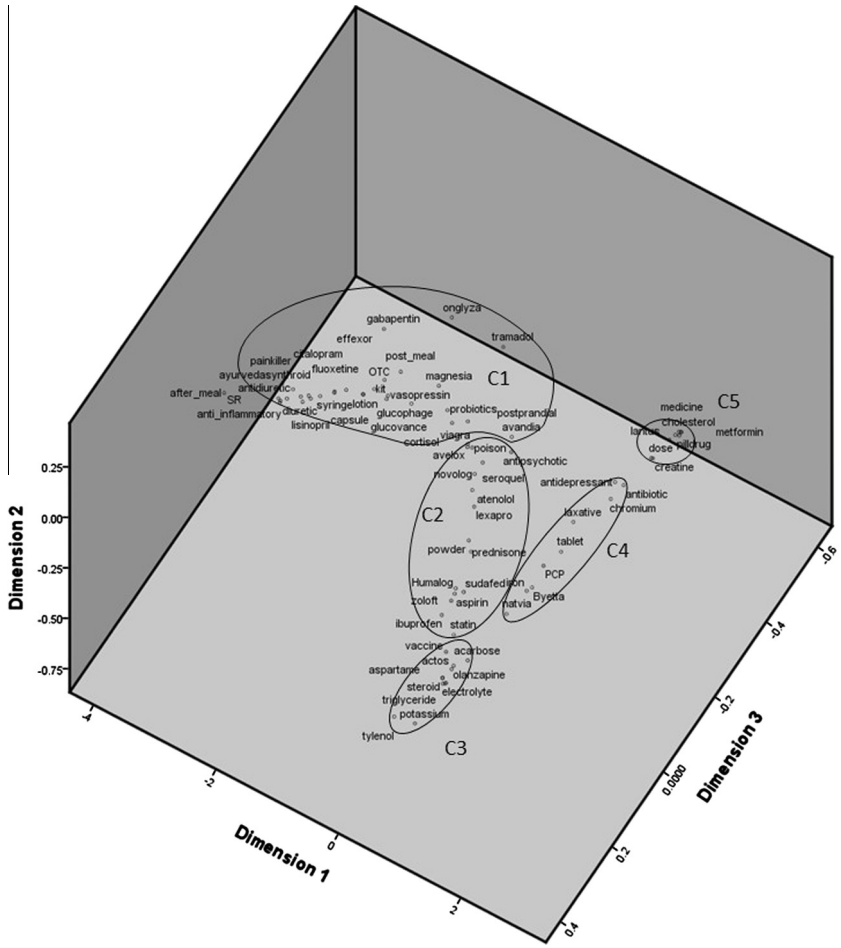

Fig. 10. The MDS display of Medication.

Table 14
Result of Medication.

Cluster	Result
C1	Anti-inflammatory, diuretic, SR, after-meal, painkiller, amoxicillin, ayurveda, antidiuretic, citalopram, lisinopril, syringe, fluoxetine, effexor, synthroid, capsule, lotion, kit, OTC, vasopressin, glucovance, glucophage, gabapentin, tramadol, postprandial, post-meal, probiotics, magnesia, cortisol, onglyza, avandia, bolus
C2	Aspirin, ibuprofen, antipsychotic, zoloft, novolog, Humalog, avelox, poison, viagra, atenolol, powder, prednisone, sudafed, statin, seroquel, lexapro
C3	Potassium, aspartame, steroid, electrolyte, triglyceride, actos, acarbose, tylenol, vaccine, olanzapine
C4	PCP, iron, natvia, Byetta, tablet, antidepressant, antibiotic, chromium, laxative
C5	Lantus, dose, creatine, pill, cholesterol, drug, metformin, medicine

Note: PCP (phencyclidine), OTC (Over-The-Counter).

4.3.1.13. Vegetable, Fruit & Grain. In the category *Vegetable*, *Fruit & Grain*, the Cosine similarity measure was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS space (Fig. 16). There were 12, 14, 12 and 38 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 20.

In this category, three distinctive clusters are identified. Cluster 4 is the fruit group (grape, cranberry, orange, apple, banana, mango, pear, melon, lemon). Cluster 2 includes the vegetable group (lettuce, green, broccoli, cauliflower, asparagus, bamboo, celery, tomato, onion), and Cluster 3 represents the grain and nuts group (pea, bakery, bean, bread, corn, grain, starch, almond, wholegrain). Cluster 1 contains a variety of recommended foods (cabbage, pumpkin, avocado, coconut, flax, berry, spinach, and cucumber) for diabetics.

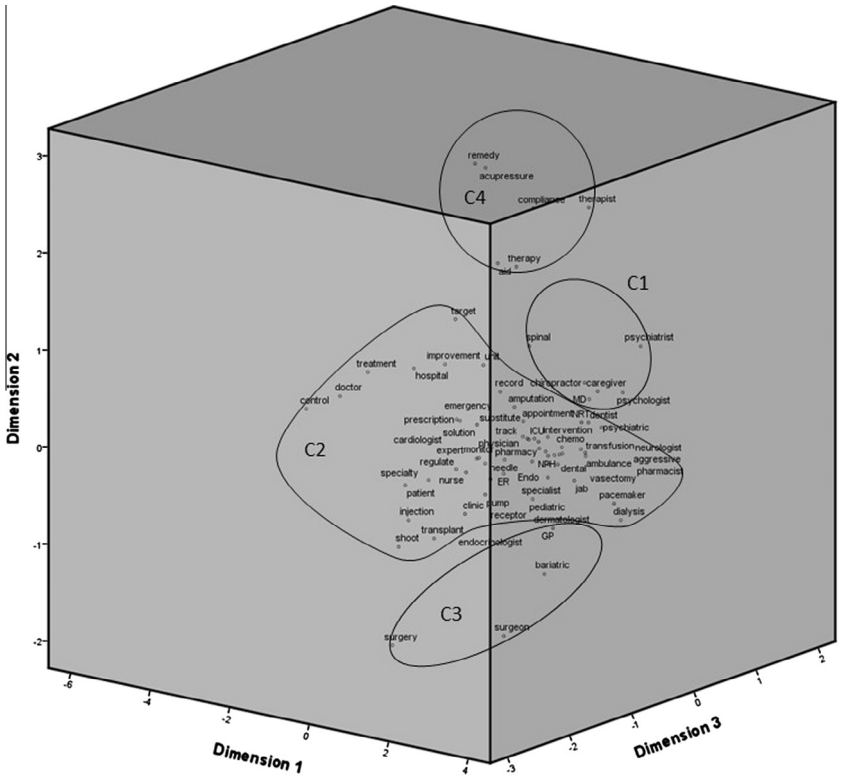

Fig. 11. The MDS display of Treatment.

Table 15
Result of Treatment.

Cluster	Result
C1	Psychologist, psychiatrist, caregiver, MD, chiropractor, spinal
C2	Pharmacy, pharmacist, NRT, aggressive, psychiatric, neurologist, dentist, dental, intervention, chemo, transfusion, vasectomy, NPH, dialysis, pacemaker, transplant, jab, specialist, clinic, emergency, ambulance, nurse, urgency, appointment, hospital, ER, monitor, physician, expert, treatment, control, doctor, prescription, improvement, patient, target, unit, substitute, solution, receptor, track, record, amputation, cardiologist, Endo, endocrinologist, pediatric, needle, ICU, injection, pump, shoot, regulate, specialty, dermatologist
C3	Bariatric, surgery, surgeon, GP
C4	Therapy, therapist, acupressure, remedy, compliance, aid

Note: NPH (Normal Pressure Hydrocephalus), NRT (Nicotine Replacement Therapy), ER (Emergency Room), ICU (Intensive-care Unit), MD (Doctor of Medicine), GP (General Practitioner).

4.3.1.14. Meat, Seafood & Dairy. In the category *Meat*, *Seafood & Dairy*, the Cosine similarity measure was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Four clusters were identified in the MDS space (Fig. 17). There were 5, 3, 16 and 3 terms in the four clusters respectively. The detailed terms of the four clusters are listed in Table 21.

Cluster 1 includes ingredients (margarine, pork, omelet, butter) that appear in a Diabetic Diet Plan for lunch. Cluster 2 includes the main ingredients (salmon, tuna, chicken) that are used in three common lunch recipes for diabetics: chicken salad, salmon salad, and tuna salad. Terms in Cluster 3 (shrimp, fish, turkey, yogurt) show up in diabetic food lists. This includes a group of the terms related to dairy (cream, milk, whey, and yogurt). Cluster 4 includes beef, marinade, and chop which are ingredients in the recipe: Barbequed Beef Kebabs with Sweet Mustard Marinade.

4.3.1.15. Fast Food, Drink & Condiment. In the category *Fast Food*, *Drink & Condiment*, the distance similarity measure ($c = 1.3$, $k = 2$) was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Three clusters were identified in the MDS space (Fig. 18). There were 11, 16 and 14 terms in the three clusters respectively. The detailed terms of the three clusters are listed in Table 22.

In general, most of the foods and drinks in this category are not healthy. As a result, they are not recommended for diabetics. Cluster 1 covers a miscellaneous collection of foods and seasonings (vinegar, Gatorade, pie, mayo) that interest

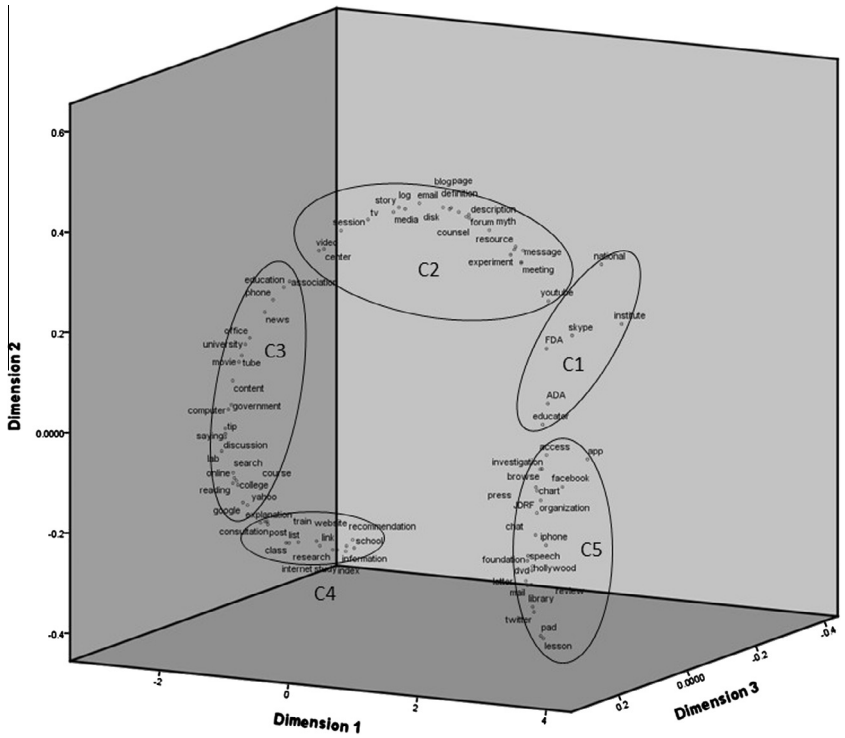

Fig. 12. The MDS display of Education & Info Resource.

Table 16
Result of Education & Info Resource.

Cluster	Result
C1	Educator, ADA, institute, national, FDA, Skype
C2	Blog, email, definition, description, myth, statistic, message, meeting, YouTube, experiment, resource, Wikipedia, counsel, forum, page, disk, media, story, log, TV, video
C3	Discussion, lab, content, university, movie, session, center, phone, association, education, news, office, tube, computer, government, saying, search, Yahoo, Google, online, tip, course, college, reading
C4	Consultation, explanation, post, train, list, class, internet, website, link, research, study, information, recommendation, index, school
C5	Lesson, pad, Twitter, library, letter, mail, review, DVD, Hollywood, foundation, speech, chat, Facebook, JDRF, organization, app, iPhone, chart, browse, press, investigation, access

Note: JDRF (Juvenile Diabetes Research Foundation).

diabetics. Cluster 2 includes substances (Pepsi, Coca Cola, honey, syrup, cookies, dessert, and ketchup) which contain a lot of sugar. Sugar increases blood glucose (blood sugar) level. Foods like popcorn, burgers, chips, and fries, also in Cluster 2, contain saturated fat. In Cluster 3, other unhealthy foods (coke, liquid, soda) are listed. It is worth pointing out that tea and coffee appear in Cluster 3 but generally are not considered unhealthy drinks.

4.3.1.16. Miscellaneous. In the category *Miscellaneous*, the Cosine similarity measure was used to create the term–term proximity matrix. The *Minkowski* power, stress value, and the *RSQ* are listed in Table 24. Three clusters were identified in the MDS space (Fig. 19). There were 13, 16, and 20 terms in the three clusters respectively. The detailed terms of the three clusters are listed in Table 23.

Cluster 1 basically suggests all diabetics should eat healthy, low-carb, and non-starchy foods at dinner, lunch, and breakfast. Cluster 2 can be characterized as foods that dietitians suggest i.e. foods rich in protein, fiber, and minerals; and low in salt, sodium, and calories. Cluster 3 includes vitamins and minerals (calcium, zinc, and vitamin B5) which are used to alleviate depression. This cluster also includes diabetes-friendly terms (gluten-free, sugar-free, low-fat).

The similarity methods used for proximity matrix generation measures in the MDS analysis, and the resulting stress values and *RSQs* for the generated categories are summarized in Table 24.

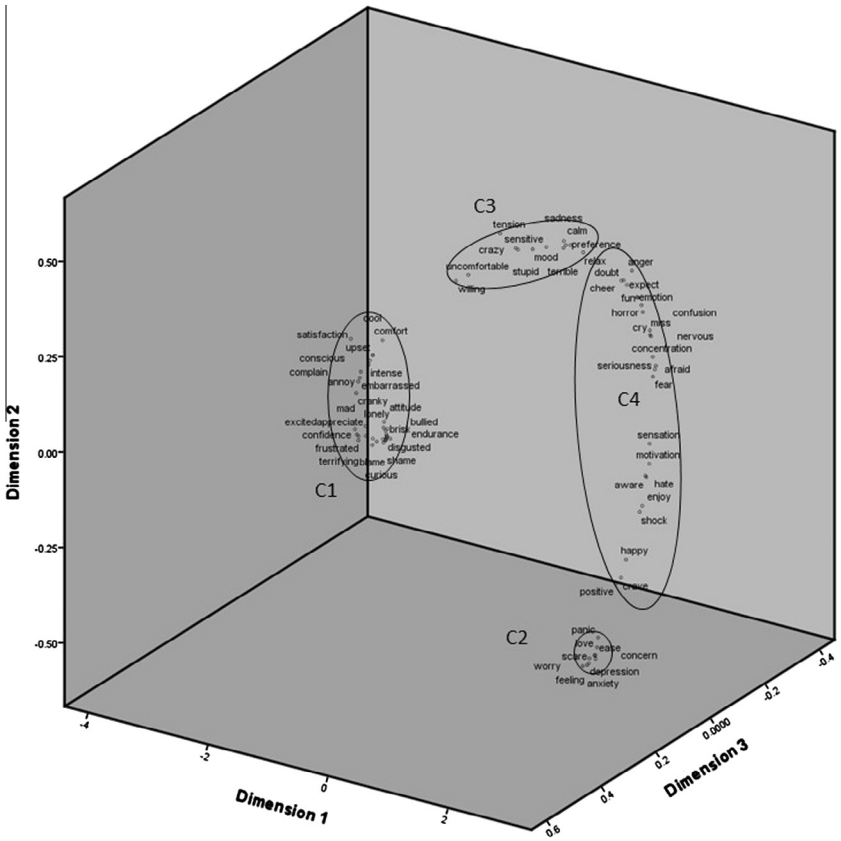

Fig. 13. The MDS display of Affect.

Table 17
Result of Affect.

Cluster	Result
C1	Excited, lonely, bullied, grateful, attitude, brisk, frightened, patience, shame, susceptible, endurance, disgusted, curious, guilt, inner, lethargy, blame, frustrated, appreciate, confidence, terrifying, mad, satisfaction, comfort, cool, complain, embarrassed, annoy, cranky, conscious, upset, intense
C2	Panic, ease, love, scare, concern, anxiety, depression, worry, feeling
C3	Willing, uncomfortable, crazy, mood, stupid, calm, relax, preference, sensitive, sadness, terrible, tension
C4	Hate, motivation, cry, miss, emotion, horror, anger, fun, doubt, expect, cheer, concentration, confusion, nervous, seriousness, afraid, fear, aware, sensation, enjoy, shock, happy, positive, crave

4.3.2. Among categories

In this section, the visualization analysis focuses on semantic relationships among the 12 identified categories. In the MDS analysis, the *Minkowski* distance measure was used and the *Minkowski* power was equal to 1. The resultant stress value was 0.02729 and the corresponding *RSQ* was equal to 0.99418. Three clusters were identified in the MDS space (Fig. 20). The detailed categories of the three clusters are listed in Table 25.

It is interesting that *Education & Info Resource*, *Affect*, *Social & Culture*, *Lifestyle*, and *Nutrient* were grouped together in Cluster 2 which is the largest cluster. These categories have a close relationship to the everyday life of diabetics, and they don't directly link to medicine. It is not surprising that *Cause & Pathophysiology*, *Diagnosis & Test*, and *Treatment* were projected together. A variety of medical professionals are included in the treatment category and it is these people who conduct medical tests, diagnose diabetes and find causes for diabetics. It is quite reasonable that in Cluster 3 *Sign & Symptom*, *Organ & Body Part*, *Complication & Related Disease*, and *Medication* are mapped onto a cluster. The category *Organ and Body Part* has a close relationship with the category *Sign & Symptom*.

4.4. Discussions

Q&A log is a user-oriented information repository. It contains more in-depth and detailed data compared to other user transaction logs such as user query transaction logs. The Q&A log includes sections, paragraphs, sentences, phrases, and

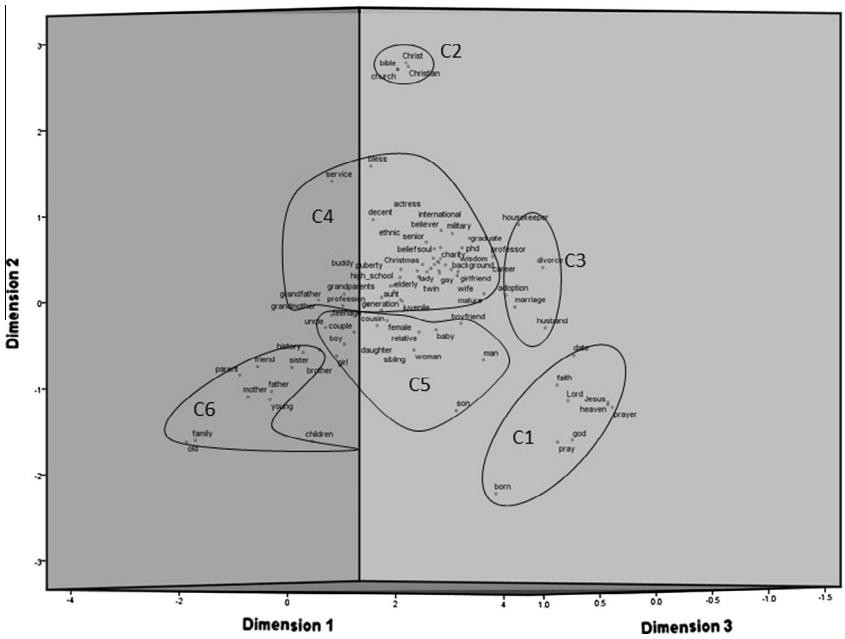

Fig. 14. The MDS display of Social & Culture.

Table 18
Result of Social & Culture.

Cluster	Result
C1	Jesus, prayer, heaven, god, pray, Lord, faith, born, date
C2	Christian, church, Christ, Bible
C3	Divorce, housekeeper, husband, marriage, adoption
C4	Believer, bless, graduate, military, profession, service, international, grandfather, senior, generation, elderly, grandmother, aunt, high-school, background, ethnic, pupil, teacher, grandparent, actress, belief, soul, puberty, twin, wisdom, charity, Christmas, gay, virgin, lady, professor, PhD, girlfriend, buddy, decent, career, mature, juvenile, wife
C5	Daughter, son, couple, female, man, woman, boyfriend, girl, teenage, boy, sibling, uncle, cousin, baby, relative
C6	Family, history, brother, sister, children, old, parent, father, mother, young, friend

terms. The user query transaction log may include sentences, phrases, and terms from users but most of queries consist of individual keywords. Using the Q&A log data to ascertain users' term uses on a specific topic of interest like diabetes should yield more data than the user query log.

For instance, the resultant data from this study show that the average number of keywords per question is 128.63, the average number of sentences per question is 8.23, the average number of keywords per response is 254.83, and the average number of sentences per response is 16.01. Twelve categories related to diabetes were yielded and analyzed respectively, the average number of valid terms among these categories is equal to 92.17, and the total number of valid terms is equal to 1106 in the investigated Q&A log. This data were greater than those found in a query transaction log study on users' term occurrence analysis of stomach, hip, stroke, depression, and cholesterol (Zhang et al., 2008), where the average number of search terms per query in the investigated transaction log was only 1.79. In another similar study (Zhang & Wolfram, Obesity-related query, 2009), the number of associated valid terms of the investigated subject obesity was equal to only 38 which is much smaller than that from the Q&A log used in this study. It implies that a Q&A log provides much more contextual information than a query transaction log. More complicated and comprehensive term relationship patterns can therefore be revealed from the Q&A log than a query transaction log.

The Q&A log comprises not only terms related to questions but also terms related to answers/responses. And the terms from both questions and answers are integrated in a record. A query transaction log includes only terms related to questions. Consequently, treatments or solutions to a symptom seldom appear in the same record as in a query transaction log. It might limit coverage of collected terms from a query transaction log. For instance, suggested special medicine for high blood sugars in pregnancy is Lantus or Levemir. A query transaction log may include the terms high blood sugars, pregnancy, and special medicine, but it is unlikely to include the responses of Lantus and Levemir.

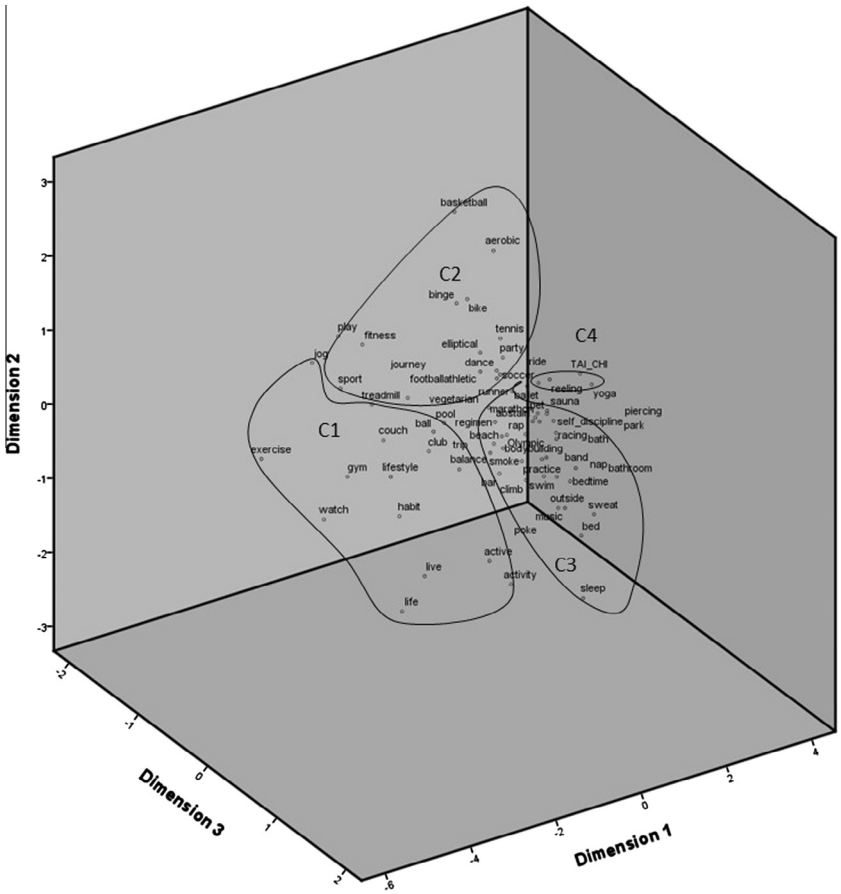

Fig. 15. The MDS display of Lifestyle.

Table 19
Result of Lifestyle.

Cluster	Result
C1	Active, activity, life, live, couch, pool, club, trip, gym, treadmill, jog, exercise, lifestyle, habit, watch, balance, ball
C2	Basketball, binge, aerobic, journey, tennis, dance, play, sport, football, soccer, bike, party, athletic, elliptical, fitness, vegetarian
C3	Bed, sleep, bedtime, music, practice, runner, band, Olympic, marathon, bodybuilding, climb, park, beach, self-discipline, bathroom, nap, piercing, racing, bath, pet, sauna, abstain, poke, ballet, sweat, swim, outside, smoke, regimen, bar, rap
C4	Tai-Chi, yoga, reeling, ride

Since the Q&A log provides in-depth context information, it reduces the ambiguity of term meaning. For instance, the term, target therapy, can be used for any incurable diseases including cancer, HIV/AIDS, Thalassemia, Muscular Dystrophy, Autism, diabetes, kidney failures, etc. However, due to a lack of context information, the meaning of this term in a keyword-based query can be ambiguous.

The implication of this study can be multifold. The generated diabetes subject schema can serve as a subject directory for existing Q&A discussion forums or subject specific portals. The clustering analysis results can be integrated into local search engines to assist user navigation. These may include newly emerging categories that reflect end-user interests and behaviors. These clustered and associated terms can also be used to enrich subject headings in thesauri and classification schemes.

The generated diabetes subject schema can serve as a subject directory for existing Q&A discussion forums like Yahoo!Answers. Currently Yahoo!Answers maintains a very simple subject directory which covers 26 very basic and general categories, and three hierarchical levels. It is not detailed enough for users to find a topic of interest like diabetes directly through browsing the hierarchy. If the produced schema is added to the subject directory, it would help not only to effectively organize questions and responses on diabetes but also to locate relevant information on diabetes. Along the same line, the user-oriented diabetes subject schema could be applied to a consumer-oriented portal related to diabetes as a subject navigation guidance system.

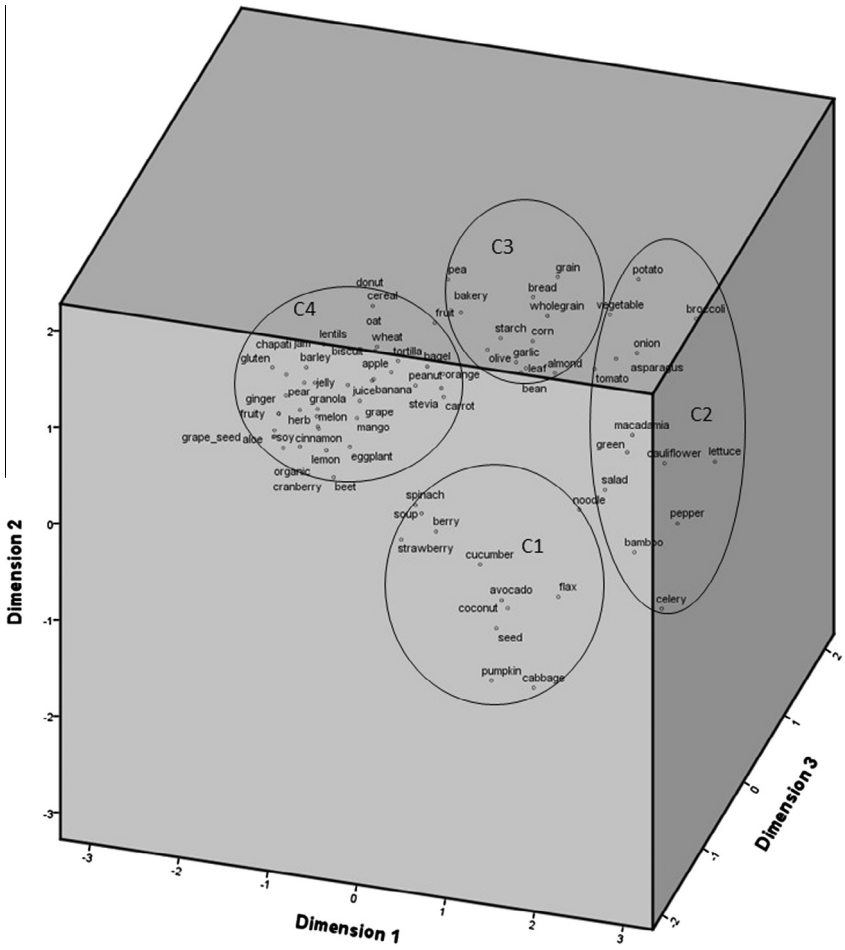

Fig. 16. The MDS display of Vegetable, Fruit & Grain.

Table 20
Result of Vegetable, Fruit & Grain.

Cluster	Result
C1	Cabbage, pumpkin, avocado, coconut, flax, seed, noodle, berry, soup, strawberry, spinach, cucumber
C2	Vegetable, salad, lettuce, green, broccoli, cauliflower, asparagus, bamboo, pepper, celery, macadamia, tomato, onion, potato
C3	Leaf, pea, bakery, bean, bread, olive, corn, grain, starch, almond, garlic, wholegrain
C4	Grape, grape-seed, soy, organic, cranberry, juice, orange, fruit, apple, banana, peanut, mango, lentils, wheat, jelly, chapatti, jam, biscuit, pear, barley, gluten, cereal, cinnamon, herb, melon, lemon, granola, fruity, ginger, aloe, eggplant, tortilla, donut, beet, oat, stevia, bagel, carrot

The clustering analysis in this study provides a group of clusters and each cluster contains a group of related terms. The clustering analysis results could also be integrated into a local search engine to enhance diabetes-related searching in a Q&A forum. When a user enters a specific search term, the search engine could prompt a list of related terms associated with that search term based on the clustering analysis results. Since this group of terms is relevant to the search term, the user could browse the list, and select other relevant terms. It could assist the user in forming his/her search strategy.

The related terms and their relationships in a cluster were generated by end-users from a Q&A log. These should reflect end user term use behaviors. The findings of this study show that health consumers very much care about the social and cultural impact of diabetes on diabetics such as the role of religion and belief in healing diabetes. End-users paid much attention to foods and drinks which can cause diabetes and affect the health of diabetics both negatively and positively such as avoiding sugar rich foods and unhealthy foods like fries and consuming healthy foods like cabbage. These searchers worried about feelings or emotional changes caused by diabetes and the potential impact on their friends and family members. These concerns included anxiety, depression, sadness, and concern of death. They were seriously concerned about the relationship between diabetics and personal life style such as participating in sport activities such as Yoga. These worries were echoed by

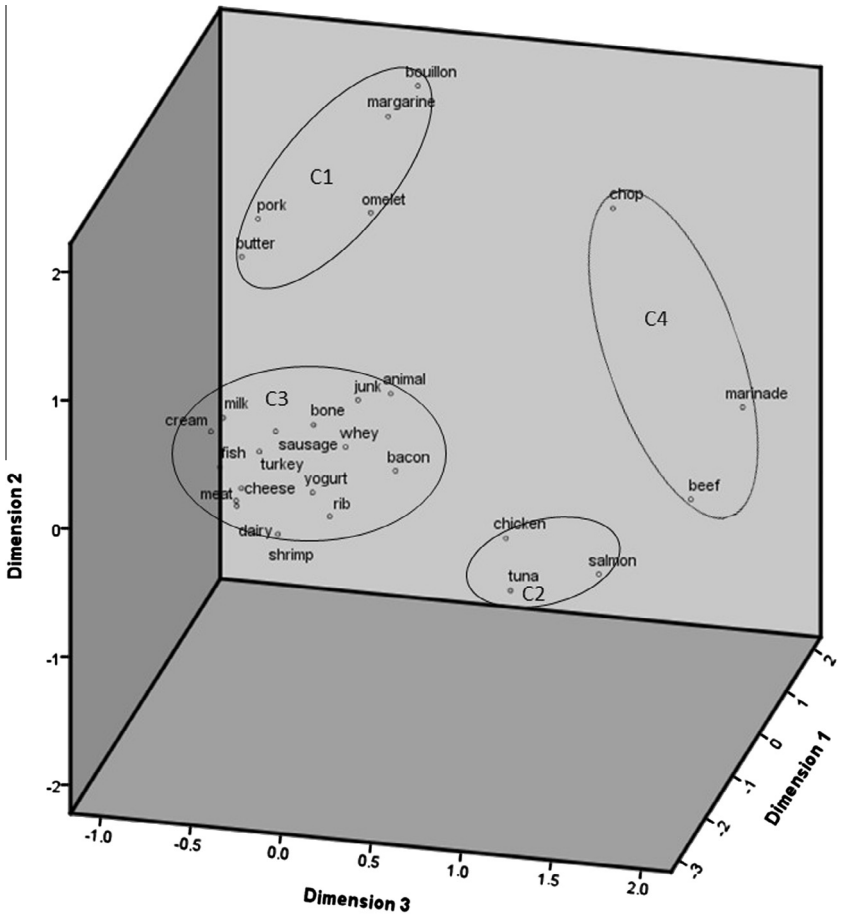

Fig. 17. The MDS display of Meat, Seafood & Dairy.

Table 21
Result of Meat, Seafood & Dairy.

Cluster	Result
C1	Bouillon, margarine, pork, omelet, butter
C2	Salmon, tuna, chicken
C3	Shrimp, cream, meat, cheese, sausage, bacon, rib, dairy, milk, fish, turkey, yogurt, whey, bone, junk, animal
C4	Beef, marinade, chop

the newly emerging categories (*Social and Culture*, *Nutrient*, *Affect*, *and Life Style*) in the schema. The high frequencies of the valid terms in these categories support this claim. The numbers of the valid terms (and corresponding percentages) of these 4 categories are 83 (7.50%), 193 (17.45%), 77 (6.96%), and 68 (6.15%) respectively. In other words, they account for about 38.06% of the investigated valid terms. It is worth pointing out that these characteristics are unique because they were not found in the existing diabetes subject schema of National Library of Medicine (PubMed: Diabetes, 2013) which is oriented to the medical professional.

Identified clusters and associated terms can also be used to enrich subject headings, thesauri and classifications such as MeSH (National Library of Medicine (NLM), MeSH, 2013), SNOMED CT (International Health Terminology Standards Development Organization, 2013), and ICD-10 (WHO, International Classification of Disease, 2013). MeSH is the controlled vocabulary thesaurus used for indexing medical articles. SNOMED CT (*Systematized Nomenclature of Medicine – Clinical Terms*) is a multilingual clinical healthcare terminology used for Electronic Health Records that contain clinical information. ICD-10 is the 10th vision of the *International Classification* (ICD) which is the standard diagnostic tool for epidemiology, health management and clinical purposes. These subject headings, thesauri and classifications are tailored to the medical professional. The cluster analysis results can provide them with not only terms but also possible term relationships from the health consumers' perspective.

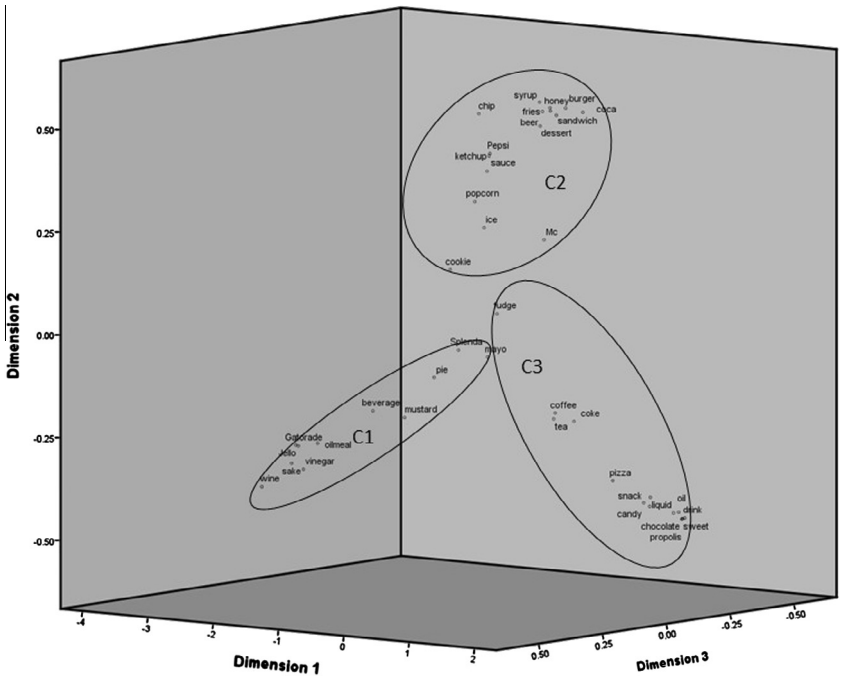

Fig. 18. The MDS display of Fast Food, Drink & Condiment.

Table 22
Result of Fast Food, Drink & Condiment.

Cluster	Result
C1	Vinegar, wine, sake, Gatorade, oilmeal, Jello, beverage, pie, mustard, mayo, Splenda
C2	Mc, burger, coca, dessert, sandwich, honey, beer, syrup, fries, cookie, chip, sauce, ketchup, Pepsi, popcorn, ice
C3	Fudge, tea, coffee, coke, pizza, snack, liquid, candy, propolis, soda, oil, sweet, chocolate, drink

One example of the usefulness of clustering analysis for this diabetes study is the grouping of user generated terms into distinct categories. For instance, in the category *Cause & Pathophysiology*, three common factors that cause diabetes (obesity, age, and gene) were identified and grouped as Cluster 1. Obesity or being overweight can result in many diseases including diabetes, arthritis, heart disease, high blood pressure, stroke, sleep apnea, and some cancers (NLM, PubMed Health: Obesity, 2013). It is recognized that age plays a role in diabetes. For instance, type I diabetes is most often diagnosed in children, teens, or young adults and type II diabetes most often occurs in adulthood (NLM, PubMed Health: Diabetes, 2013). Some people are more likely genetically to get diabetes than others. For instance, type II diabetes has a strong link to family history (ADA, Genetics of Diabetes, 2013). Similarly, cluster analysis of the terms in the category *Sign & Symptom* identified and grouped 8 of 10 diabetes symptoms posted in the Medical Encyclopedia (ADA, Diabetes Basics, 2013) in a cluster labeled Cluster 4. These included blurry vision (blurry, vision, sight), unusual thirsty (excessive, thirsty, polydipsia), extreme hunger(hunger), cuts/bruises are slow to heal (slow-healing, heal, slow, bruise, wound), frequent infections (frequency, infection), tingling/numbness in the hands/feet (tingle, numbness), recurring skin, gum, or bladder infections (skin, infection, itch), and frequent urination (frequency, urination). The terms in parentheses were user-generated terms that collected in Cluster 4 of *Sign & Symptom*. It is interesting that the term palpitation appears in Cluster 4 but it is not related to any of the 10 diabetes symptoms. The two remaining diabetes symptoms, not included in the cluster are unusual weight loss and extreme fatigue. But Cluster 3 includes the terms fatigue, tiredness, and exhaustion which are related to extreme fatigue. A third and final example demonstrates how clustering analysis mirrors known groupings of terms and is demonstrated in the subcategory *Vegetable*, *Fruit & Grain*, where the fruit-related terms (grape, grape-seed, cranberry, juice, orange, fruit, apple, banana, mango, pear, melon, fruity) were clustered together in Cluster 4. All of these except lemon were found on the fruit list recommended by American Diabetes Association (MyFoodAdvisor, 2013).

The Q&A transaction log analysis, however, is not without problems. It is apparent that the visual clustering analyses in this study were conducted based on free-texts in the Q&A log. Compared with a query in a query transaction log, the positive side of a Q&A log is that the dialog of questions and answers provides researchers with a large amount of rich textual data in each record. However, the downside of the Q&A log is that not every individual term extracted from a Q&A record is highly

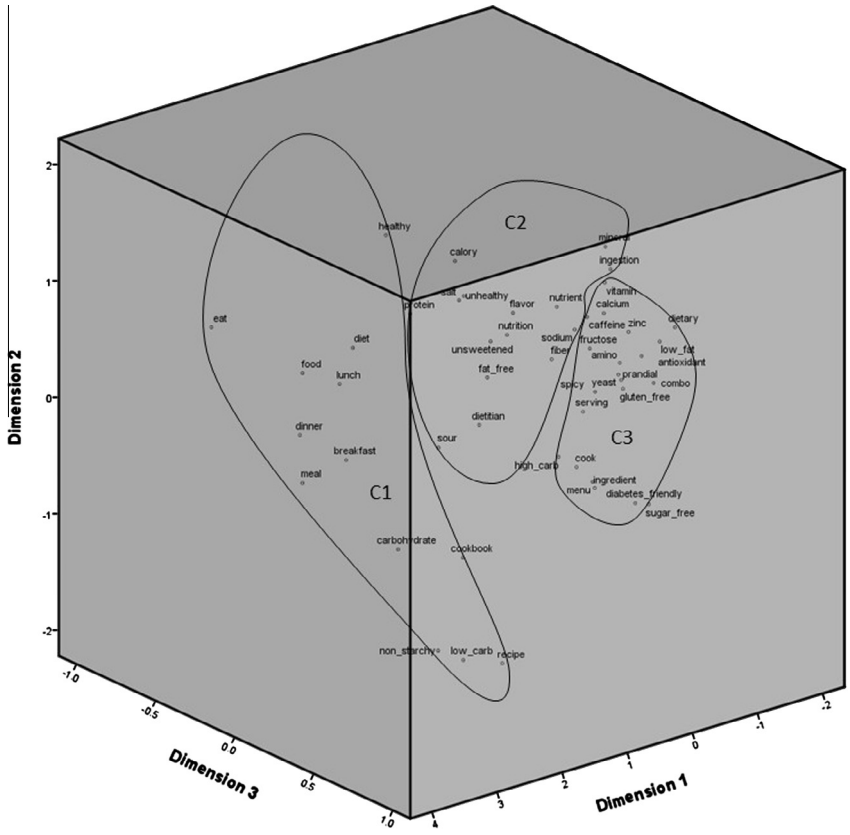

Fig. 19. The MDS display of Miscellaneous.

Table 23
Result of Miscellaneous.

Cluster	Result
C1	Eat, food, diet, healthy, carbohydrate, meal, dinner, lunch, breakfast, low-carb, non-starchy, recipe, cookbook
C2	Sour, salt, sodium, cook, fat-free, unhealthy, nutrition, calorie, nutrient, protein, fiber, dietitian, mineral, unsweetened, flavor, ingestion
C3	Calcium, zinc, caffeine, yeast, combo, prandial, gluten-free, amino, menu, sugar-free, diabetes-friendly, fructose, ingredient, high-carb, low-fat, serving, vitamin, dietary, antioxidant, spicy

relevant to the topic of interest. In a question and answer dialog the responses are uncontrolled and can be quite open. These natural language sentences may consist of many terms irrelevant to the topic. As the number of different words in a Q&A record increases, the chance that a low relevant word is included in the record increases. As a result, a cluster may not have a clear theme, or terms in the cluster based on a Q&A log may not be semantically related to each other such as in Cluster 3 of the category of *Sign and Symptom* (see Fig. 5) and Cluster 2 of the category of *Organs and Body Parts* (see Fig. 8). It is a weakness of a Q&A log analysis.

It is also important to note that some types of questions tend to receive more responses than other types. The number of responses to a question tended to increase if the question was well stated, detailed, specific, and/or complex. For instance, a consumer posted a question which included his age, gender, medical history, medical test results, health condition, psychological and mental situation, working condition, social environment, and symptoms. It resulted in 8 relevant responses. This is considered a large number of responses in this Q&A log. In addition, a question with a sincere tone also tended to attract more responses.

It is no surprise that both words and sentences in questions outnumbered words and sentences in responses significantly in the investigated *Q&A* data set. That is because a question is usually created by an individual questioner while responses to the question can be produced by multiple people. In addition, answers to a question can be addressed from multiple perspectives.

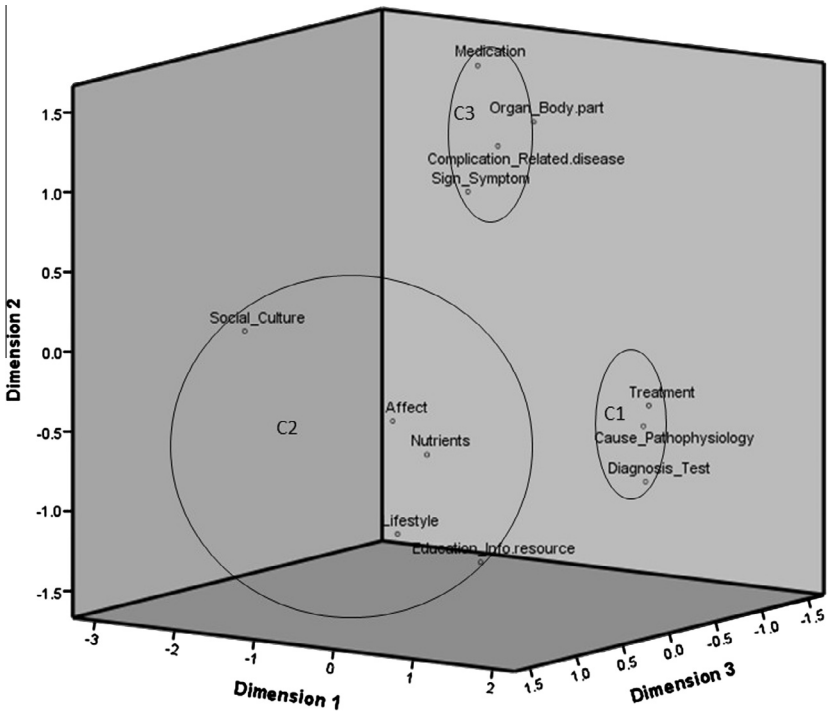

Fig. 20. The MDS display of 12 categories.

Table 24
Result summary of the term analysis within a category.

Category	The similarity used for proximity	Measure in MDS	Stress value	RSQ
1. Cause & Pathophysiology	Cos	Minkowski power = 1	0.10238	0.94986
2. Sign & Symptom	Cos	Minkowski power = 1	0.09232	0.96366
3.1. Diagnosis	Cos	Minkowski power = 1	0.06797	0.98444
3.2. Test	Cos	Minkowski power = 1	0.07698	0.98296
4. Organ & Body Part	Distance $c = 1.3$, $k = 2$	Minkowski power = 1	0.01399	0.99944
5. Complication & Related Disease	Distance $c = 1.3$, $k = 2$	Minkowski power = 1	0.01237	0.99949
6. Medication	Distance $c = 1.3$, $k = 2$	Minkowski power = 1	0.01384	0.99939
7. Treatment	Cos	Minkowski power = 1	0.11390	0.94937
8. Education & Info Resource	Distance $c = 1.3$, $k = 2$	Minkowski power = 1	0.01183	0.99956
9. Affect	Distance $c = 1.3$, $k = 2$	Minkowski power = 1	0.1402	0.99940
10. Social & Culture	Cos	Minkowski power = 1	0.08539	0.97796
11. Lifestyle	Cos	Minkowski power = 1	0.11641	0.94433
12.1. Vegetable, Fruit & Grain	Cos	Minkowski power = 1	0.1014	0.94453
12.2. Meat, Seafood & Dairy	Cos	Minkowski power = 1	0.09779	0.93132
12.3. Fast Food, Drink & Condiment	Distance $c = 1.3$, $k = 2$	Minkowski power = 1	0.02579	0.99795
12.4. Miscellaneous	Cos	Minkowski power = 1	0.07947	0.97486

Table 25
Display of the three clusters in the category analysis.

Cluster	Result
C1	Cause_Pathophysiology, Diagnosis_Test, Treatment
C2	Education_Info.resource, Affect, Social_Culture, Lifestyle, Nutrient
C3	Sign_Symptom, Organ_Body.part, Complication_Related.disease, Medication

The coding analysis was conducted by the researchers and the medical professional respectively. An inter-coder reliability analysis was also conducted by using the statistical Cohen's Kappa method. The corresponding Cohen's Kappa coefficient for this inter-coder reliability analysis is 0.792.

5. Conclusion

Consumer health informatics is a rapidly growing field. It addresses health related issues from the consumer or patient view rather than the medical professional view. As more and more health information resources become available online and Web 2.0 technology matures, health consumers will become more active in online information access and consultation. The social Q&A forums like Yahoo!Answers will continue to attract more health consumers and patients. It is natural to utilize user Q&A logs to study health consumers' behaviors and to reveal important patterns.

Diabetes is a chronic disease. It can cause high levels of sugar in the patient's blood resulting in a variety of complications and even death. Diabetes is a disease among all ethnic groups, gender groups, and age groups. Therefore, a study on users' diabetes information seeking patterns is important.

In this study, the social Q&A log in Yahoo!Answers was used. There were 2565 valid records extracted from Yahoo!Answers. The total number of extracted words from the records was 1,043,158. The average number of words per question was 128.63; average number of sentences per question was 8.23; average number of words per response was 254.83; and average number of sentences per response was 16.01 in the diabetes related Q&A records. The newly emerged schema from the Q&A log analysis had 12 categories: *Cause & Pathophysiology*, *Sign & Symptom*, *Diagnosis & Test*, *Organ & Body Part*, *Complication & Related Disease*, *Medication*, *Treatment*, *Education & Info Resource*, *Affect*, *Social & Culture*, *Lifestyle*, and *Nutrient.* It is no surprise that the category *Nutrient* clenched the first position and accounting for 17.31% of the terms; *Diagnosis & Test* held the second position (for 11.39% of the terms); and *Complication & Related Disease* maintained the third position (for 8.88% of the terms).

The visualization analysis on diabetes related terms from the Q&A log was conducted based on the emerging diabetes schema at two levels: term analysis within categories where semantic relationships among terms within a category were analyzed; and category analysis within the schema where the semantic relationships among the 12 categories were revealed. The analyses at the two levels show that terms and categories were clustered and patterns were revealed.

Future research directions on this topic include, but are not limited to, term analysis between two closely related categories. As shown, there are close relationships among the emerging categories. If two closely related categories such as the *Sign & Symptom* category and *Organ & Body Part* category; the *Diagnosis* category and *Test* category; or the *Test* category and *Treatment* category are selected for a cross category analysis, it would reveal more interesting findings related to diabetes from a quite different angle. In this study, one of the primary focuses was to conduct a term clustering analysis within a category. In other words, term semantic relationships were revealed and examined only for terms within a category, and semantic relationships among terms from two different categories were excluded. In fact, terms from two different categories may maintain close relationships. For instance, terms from the *Diagnosis* category may be related to terms from the *Sign & Symptom* category. If a visual clustering analysis were conducted, the analysis may unveil interesting findings such as the connection between a certain symptom and a specific type of diabetes. Another research direction will be to apply the same research methodology to other health topics.

References

Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008). Knowledge sharing and Yahoo Answers: Everyone knows something. In *Proceeding of the 17th international conference on World Wide Web (WWW '08)* (pp. 665–674). New York: ACM.

Alexa (2012). *The top ranked sites in references category*. <http://www.alexa.com/topsites/category/Top/Reference> Accessed 18.04.12.

American Diabetes Association (2012). *Diabetes statistics*. <http://www.diabetes.org/diabetes-basics/diabetes-statistics/?loc=DropDownDB-stats> Accessed 20.04.12.

American Diabetes Association (2013). *Diabetes basics: Symptoms*. <http://www.diabetes.org/diabetes-basics/symptoms/> Accessed 18.02.13.

American Diabetes Association (2013). *Genetics of diabetes*. <http://www.diabetes.org/diabetes-basics/genetics-of-diabetes.html> Accessed 18.02.13.

American Diabetes Association (2013). *MyFoodAdvisor*. <http://tracker.diabetes.org/explore/browse/fruit/?page=1> Accessed 18.02.13.

Brink, S. J., Miller & Moltz, K. C. (2002). Education and multidisciplinary team care concepts for pediatric and adolescent diabetes mellitus. *J. Pediatric Endocrinology & Metabolism, 15*(8), 1113–1130.

Brink, S. J., & Chiarelli, F. G. (2004). Education and multidisciplinary team approach in childhood diabetes. *Acta Bio Medica, 75*(1), 7–21.

Centers for Disease Control and Prevention (2012). *10 Leading causes of death by age group, United States – 2009*. Atlanta, GA: U.S. National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Eerola, J., & Vakkari, P. (2008). How a general and a specific thesaurus cover expressions in patients' questions and physicians' answers. *Journal of Documentation, 64*(1), 131–142.

Fox, C. (1989). A stop list for general text. *SIGIR Forum, 24*(1–2), 19–21.

Fox, S. (2006). *Online health search 2006*. Washington, DC: Pew Internet & American Life Project. Washington, DC: Pew Research Center. <http://www.pewinternet.org/Reports/2006/Online-Health-Search-2006.aspx> Accessed 16.01.12.

Fox, S. (2008). *The engaged e-patient population*. Pew Internet & American Life Project. Washington, DC: Pew Research Center. <http://nicolaziady.com/wp-content/uploads/2011/07/PIP_Health_Aug08.pdf.pdf> Accessed 16.01.12.

Gazan, R. (2011). Social Q&A. *Journal of the American Society for Information Science and Technology, 62*(12), 2301–2312.

Gazan, R. (2006). Specialists and synthesists in a question answering community. In *Proceedings of the 69th annual meeting of the American society for information science & technology (ASIST'06)* (pp. 1–10). Medford, NJ: Information Today.

Gooden, R. J., & Winefield, H. R. (2007). Breast and prostate cancer online discussion boards—A thematic analysis of gender differences and similarities. *Journal of Health Psychology, 12*(1), 103–114.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*, 255–274.
Hakanen, E. A., & Wolfram, D. (1995). Citation relationships among international mass communication journals. *Journal of Information Science, 22*(3), 9–15.
Harper, F., Moy, D., & Konstan, J. (2009). Facts or friends? Distinguishing informational and conversational questions in social Q&A sites. In *Proceedings of the 27th international conference on human factors in computing systems (CHI 2009)* (pp. 759–768). New York, NY: ACM.
Harper, F., Raban, D., Rafaeli, S., & Konstan, J. (2008). Predictors of answer quality in online Q&A sites. In *Proceedings of the 26th annual SIGCHI conference on human factors in computing systems* (pp. 865–874). New York: ACM.
Hesse, B. W., Nelson, D. E., Kreps, G. L., Croyle, R. T., Arora, N., Rimer, B. K., et al (2005). The impact of the internet and its implications for health care providers: Findings from the first Health Information National Trends Survey. *Archives of Internal Medicine, 165*(22), 2618–2624.
Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. New York: The Guilford Press.
International Health Terminology Standards Development Organization (2013). *SNOMED CT*. <http://www.ihtsdo.org/snomed-ct/> Accessed 11.02.13.
Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In *Proceedings of the 16th ACM conference on information and knowledge management (CIKM ' 07)* (pp. 919–22), New York, NY: ACM.
Kelly, D., Wacholder, N., Rittman, R., Sun, Y., Kantor, P., Small, S., et al (2007). Using interview data to identify evaluation criteria for interactive, question-answering systems. *Journal of the American Society for Information Science and Technology, 58*(7), 1032–1043.
Keselman, A., Browne, A., & Kaufman, D. (2008). Consumer health information seeking as hypothesis testing. *Journal of the American Medical Informatics Association, 15*(4), 484–495.
Kim, S., & Oh, S. (2009). Users' relevance criteria for evaluating answers in a social Q&A site. *Journal of the American Society for Information Science and Technology, 60*, 716–727.
Kim, S., Oh, J., & Oh, S. (2007). Best-answer selection criteria in a social Q&A site from the user-centered relevance perspective. In *Proceedings of the 70th annual meeting of the American society for information science and technology (ASIST '07)* (pp. 1–15). Medford, NJ: Information Today.
Kim, S., Oh, S. & Oh, J. (2008). Evaluating health answers in a social Q&A site. In *Proceedings of the American society for information science and technology (ASIST'08)* (pp. 1–6). Columbus, Ohio: Information Today.
Longo, D., Schubert, S., Wright, B., LeMaster, J., Williams, C., & Clore, J. (2010). Health information seeking, receipt, and use in diabetes self-management. *Annals of Family Medicine, 8*(4), 334–340.
McCray, A., & Tse, T. (2003). Understanding search failures in consumer health information systems. In *AMIA annual symposium proceedings* (pp. 430–434). Bethesda, MD: National Library of Medicine.
Milewski, J., & Chen, Y. (2010). Barriers of obtaining health information among diabetes patients. *Studies in Health Technology and Informatics, 160*(1), 18–22.
National Library of Medicine (2013). *MeSH*. <http://www.ncbi.nlm.nih.gov/mesh> Accessed 11.02.13.
National Library of Medicine (2013). *Pub Med Health: Diabetes*. <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0002194/> Accessed 18.02.13.
National Library of Medicine (2013). *PubMed Health: Obesity*. <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0004552/> Accessed 18.02.13.
Nordfeldt, S., Hanberger, L., & Bertero, C. (2010). Patient and parent views on a web 2.0 diabetes portal—The management tool, the generator, and the gatekeeper: Qualitative study. *Journal of Medical Internet Research, 12*(2), e17.
Nordfeldt, S., Johansson, C., Carlsson, E., & Hammersjo, J. A. (2005). Use of the Internet to search for information in type 1 diabetes children and adolescents: A cross-sectional study. *Technology and Health Care, 13*(1), 67–74.
Poikonen, T., & Vakkari, P. (2009). Lay persons' and professionals' nutrition-related vocabularies and their matching to a general and a specific thesaurus. *Journal of Information Science, 35*(2), 232–243.
Raban, D. R. (2009). Self-presentation and the value of information on Q&A sites. *Journal of the American Society for Information Science and Technology, 60*(12), 2465–2473.
Rosenbaum, H., & Shachaf, P. (2010). A structuration approach to online communities of practice: The case of Q&A communities. *Journal of the American Society for Information Science and Technology, 61*(9), 1933–1944.
Roush, W. (2006). *What's the best Q&A site?* Technology Review. <http://www.technologyreview.com/communications/17932/?a=f> Accessed 16.11.11.
Seidman, J., Steinwachs, D., & Rubin, H. (2003). Conceptual framework for a new tool for evaluating the quality of diabetes consumer-information web sites. *Journal of Medical Internet Research, 5*(4), e29.
Shah, C., Oh, J. S., & Oh, S. (2008). Exploring characteristics and effects of user participation in online social Q&A sites. *First Monday*, *13*(9). <http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2182/2028> Accessed 15.01.12.
Shah, C., Oh, S., & Oh, J. S. (2009). Research agenda for social Q&A. *Library & Information Science Research, 31*(4), 205–209.
Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mapping. *Journal of Information Science, 11*, 147–159.
Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., Ozmutlu, S., et al (2004). A study of medical and health queries to web search engines. *Health Information and Libraries Journal, 21*, 44–51.
Stvilia, B., Mon, L., & Yi, Y. (2009). A model for online consumer health information quality. *Journal of American Society of Information Science and Technology, 60*(9), 1781–1791.
Thelwall, M. (2002). An initial exploration of the link relationship between UK university Web sites. *ASLIB Proceedings, 54*(2), 118–126.
Vaughan, L. (2006). Visualizing linguistic and cultural differences using Web co-link data. *Journal of the American Society for Information Science and Technology, 57*(9), 1178–1193.
White, R. W., Dumais, S., & Teevan, J. (2008). How medical expertise influences web search interaction. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'08)* (pp. 791–792). New York, NY: ACM.
World Health Organization (2012). *Diabetes*. <http://www.who.int/mediacentre/factsheets/fs312/en/> Accessed 20.01.12.
World Health Organization (2013). *International classification of diseases (ICD)*. <http://www.who.int/classifications/icd/en/> Accessed 11.02.13.
Yahoo!Answers (2012). *All categories*. <http://answers.yahoo.com/> Retrieved 01.11.11.
Yoon, J., & Chung, E. (2011). Understanding image needs in daily life by analyzing question in a social Q&A site. *Journal of the American Society for Information Science and Technology, 62*(11), 2201–2213.
York, J., Bohn, S., Pennock, K., & Lantrip, D. (1995). Clustering and dimensionality reduction in SPIRE. In AIPA Steering Group (Ed.), *Proceedings of the symposium on advanced intelligence processing and analysis* (pp. 73). Washington, DC: Office of Research and Development.
Zeng, Q. T., Kogan, S., Ash, N., Greenes, R. A., & Boxwala, A. A. (2002). Characteristics of consumer terminology for health information retrieval. *Methods of Information in Medicine, 41*, 289–298.
Zeng, Q. T., Tse, T., Crowell, J., Divita, G., Roth, L., & Browne, A. C. (2005). Identifying consumer-friendly display (CFD) names for health concepts. In C. P. Friedman et al. (Eds.), *AMIA symposium proceedings* (pp. 859–863). Washington, DC: AMIA.
Zhang, J. (2008). *Visualization for information retrieval*. Berlin, Heidelberg: Springer.
Zhang, J., & Rasmussen, E. (2001). Developing a new similarity measure from two different perspectives. *Information Processing & Management, 37*(2), 279–294.
Zhang, J., & Wolfram, D. (2001). Visualization of term discrimination analysis. *Journal of the American Society for Information Science and Technology, 52*(8), 615–627.
Zhang, J., & Wolfram, D. (2009). Visual analysis of obesity-related query terms on HealthLink. *Online Information Review, 33*(1), 43–57.
Zhang, J., Wolfram, D., Wang, P., Hong, Y., & Gillis, R. (2008). Visualization of health-subject analysis based on query term co-occurrences. *Journal of the American Society for Information Science and Technology, 59*(12), 1933–1947.
Zhang, Y. (2010). Contextualizing consumer health information searching: An analysis of questions in a social Q&A community. In *Proceedings of the 1st ACM international health informatics symposium* (pp. 210–219). New York: ACM.