Analysis of Query Keywords of Sports-Related Queries Using Visualization and Clustering

Jin Zhang and Dietmar Wolfram
School of Information Studies, University of Wisconsin—Milwaukee, Milwaukee, WI 53201.
E-mail: {jzhang, dwolfram}@uwm.edu

Peiling Wang
School of Information Sciences, College of Communication and Information,
University of Tennessee at Knoxville, Knoxville, TN 37996–0341. E-mail: peilingw@utk.edu

The authors investigated 11 sports-related query keywords extracted from a public search engine query log to better understand sports-related information seeking on the Internet. After the query log contents were cleaned and query data were parsed, popular sports-related keywords were identified, along with frequently co-occurring query terms associated with the identified keywords. Relationships among each sports-related focus keyword and its related keywords were characterized and grouped using multidimensional scaling (MDS) in combination with traditional hierarchical clustering methods. The two approaches were synthesized in a visual context by highlighting the results of the hierarchical clustering analysis in the visual MDS configuration. Important events, people, subjects, merchandise, and so on related to a sport were illustrated, and relationships among the sports were analyzed. A small-scale comparative study of sports searches with and without term assistance was conducted. Searches that used search term assistance by relying on previous query term relationships outperformed the searches without the search term assistance. The findings of this study provide insights into sports information seeking behavior on the Internet. The developed method also may be applied to other query log subject areas.

Introduction and Previous Research

Transaction log data analysis has been widely used over the past few decades to better understand user searching. The reasons for this are clear. Large quantities of search information are faithfully recorded in a transaction log. It is, therefore, natural to tap into transaction logs to gain insights into users' search behavior. A log file is able to record the history of all online users' requests. It accurately keeps and maintains all online users' activities performed on a server. Usually, the user's request includes the client Internet Protocol (IP) address, request date/time, page requested, HTTP code, bytes served, user agent, referrer, and so on. The data are kept in a standard format in a transaction log file (Hallam-Baker & Behlendorf, 2008).

Although a transaction log comprises rich data, including browsing times and traversal paths, it is the queries directly submitted by users that have attracted the most research attention. Query data contain keywords that reflect users' wide-ranging information needs. Query logs have been analyzed from a variety of sources with different emphases and audiences. Studies of query characteristics have included special-topic Web sites such as *THOMAS* (Croft, Cook, & Wilder, 1995), a local search engine (Park, Lee, & Bae, 2005), a university Web site (Wang, Berry, & Yang, 2003), digital libraries (Jones, Cunningham, & McNab, 1998; Mahoui & Cunningham, 2000), a Web-based, online public-access catalog (Cooper, 2001), and bibliographic databases (Yi, Beheshti, Cole, Leide, & Large, 2006) as well as public search engines such as Fireball (Hoelscher, 1998), AltaVista (Jansen, Jansen, & Spink, 2005), Excite (Jansen, Goodrum, & Spink, 2000; Rieh & Xie, 2006; Silverstein, Marais, Henzinger, & Moricz, 1999; Spink, Jansen, Wolfram, & Saracevic, 2002; Spink, Wolfram, Jansen, & Saracevic, 2001), and Vivisimo (Koshman, Spink, & Jansen, 2006) and federated search systems such as Dogpile (Spink, Jansen, & Koshman, 2007). Similarly, a number of studies have focused on specific search areas or circumstances. These have included search queries on mobile phones in Japan (Baeza-Yates, Dupret, & Velasco, 2007), employment-based searching (Jansen et al., 2005), multimedia searching on a public search engine (Goodrum & Spink, 2001; Jansen et al., 2000), sexual topics (Spink, Ozmutlu, & Lorence, 2004), and health or medical-related information (Spink, Yang, et al., 2004).

Different approaches have been applied to the analysis and reporting of query data. Early studies such as those

Received May 21, 2008; revised March 15, 2009; accepted March 17, 2009

 DOI: 10.1002/asi.21098

by Croft et al. (1995), Hoelscher (1998), and Silverstein et al. (1999) have provided descriptive analyses of keywords, queries, and possibly, sessions. Ross and Wolfram (2000) used hierarchical cluster analysis on query-keyword pairs for multi-keyword queries for a transaction log from the Excite search engine to identify groups of topics. Beitzel, Jensen, Chowdhury, Frieder, & Grossman (2007) argued that categorization and classification of user queries can lead to increased effectiveness and efficiency in general-purpose Web search systems. They investigated properties of a very large query log over varying periods and were able to identify and examine topical trends. Shi and Yang (2007) developed a method to assist users in formulating initial queries using association rules to identify related queries from a Web log. In a similar study, Huang, Chien, and Oyang (2003) were able to suggest keywords for assisting user interactive searches. Relevant keywords suggested for a user query came from those that co-occurred in similar query sessions in a transaction log. Beeferman and Berger (2000) applied an agglomerative clustering algorithm to search queries and relevant Web pages in a transaction log to discover potential query clusters. Likewise, Wen, Nie, and Zhang (2001) studied the relationship between a search query and selected Web pages in conjunction with query contents to categorize queries in a transaction log. To visually analyze Web logs, Whittle, Eaglestone, Ford, Gillet, and Madden (2007) looked for similarities between queries and identified sequences of "query transformations," which were represented as graphical networks, thereby providing a different view of search behavior. To facilitate transaction log analysis, Joshi, Joshi, Yesha, and Krishnapuram (1999) developed an ad hoc tool for analytic queries on a transaction log warehouse.

Sports-related information-seeking behavior on the Internet has attracted the attention of several researchers. Ernest, Level, and Culbertson (2005) concluded that the Internet was a preferred source for sports and other consumer information. Sports topics were identified as one of the most popular online Internet search-topic categories by Kelly (2006). The role that the Internet plays for sports fans is a direct extension of their interest in sports. People who like to watch sports are more likely to purchase event tickets and visit sports-related Web sites (SportsBusiness Daily, 2001). Due to the nature of sports, many studies of the sports-related information-seeking behavior have focused on children and young people. The Pew Internet and American Life Project (2005) reported that online entertainment information seeking, such as sports, remained the most popular use of the Internet for teenage users. A recent Oxford Internet Survey found that young people are more likely to seek online information about sports, humorous content, and employment than are working or retired adults, but they are less likely to look for information pertaining to health, medical care, or travel (Dutton & Helpser, 2007). In another study on children's online information-seeking behavior (Hirsh, 1999), sports-related topics were used as search tasks in the study. Participants employed a variety of information sources to search the sports-related topics. Children's search processes were observed and analyzed, and their information search behavior patterns were generalized. Based on these studies, it is clear that although several studies have examined sports-related information on the Internet, few have focused on sports-related information-seeking behavior.

Existing studies have demonstrated the complexity of relationships in how vocabularies are used to formulate queries. Clustering or grouping techniques lend themselves to newer exploratory methods such as information visualization. Information visualization was introduced to utilize the human perceptional capacity to present, understand, and explore complex abstract information by using computing techniques (Robertson, Card, & Mackinlay, 1989). It is widely recognized that the human perception system is responsible for not only receiving outside information but also processing received information. The way information is processed by the human perception system is unique, effective, and efficient. According to Zeki (1992), the four parallel systems within the human visual cortex work simultaneously to process received visual inputs from the retina.

Information visualization is employed to reveal connections and relationships among investigated objects. It is believed that information visualization can be used to support tasks such as data analysis, information exploration, information explanation, trend prediction, and pattern detection (Zhang, 2008).

A traditional clustering analysis usually offers a set of separate and disconnected clusters. A visual-analysis method provides users with proximity characteristics of objects that can be visually grouped into clusters; the connections between an object and multiple relevant objects in a cluster; a holistic overview of all involved objects and clusters; and the contexts and degree to which clusters and objects are connected and related. In addition, a visual display of clusters is vivid, straightforward, and intuitive.

There are many available information-visualization techniques and applications such as Pathfinder associative networks (Fowler, Fowler, & Wilson, 1991; Schvaneveldt, Durso, & Dearholt, 1989), which are appropriate for visualizing a simplified and optimized network for a sophisticated network; self-organizing maps (*SOM*s; Kohonen, 2001; Kohonen et al., 2000), which categorize objects by presenting them in a semantic map; Euclidean spatial-based visualization approaches (Zhang, 2001; Zhang & Korfhage, 1999), which employ reference points to create a customized visual configuration for users; and multidimensional scaling (MDS) analysis, to name just a few.

Multidimensional scaling was first introduced by Torgerson (1952) and was further developed by Kruskal (1964a, 1964b). Since the MDS method can effectively reveal hidden relationships among objects based on their proximities, it has been applied to co-citation analysis, where the proximity between two objects can be clearly defined (White & McCain, 1998). If a proximity relationship is defined differently, the MDS visual analysis can be made at different levels for various objects. When the proximity relationship between two objects is defined as the number of

DOI: 10.1002/asi

references that two documents co-cite, then the displayed objects are documents, and the visual analysis stays at a document level (Chalmers & Chitson, 1992; York, Bohn, Pennock, & Lantrip, 1995). When the proximity relationship between two objects is described as the number of the journals that two journals co-cite, then the visualized objects in the visual space are journals, and the visual analysis stays at a higher journal level (Hakanen & Wolfram, 1995). However, if the proximity relationship between two objects is defined as the number of research areas that two areas co-cite in a database, then the visualized objects in the visual space are research areas, and the visual analysis remains at a research area level (Small, 1973; Small & Garfield, 1985). It is no surprise that the MDS method also has been applied to Internet information analysis. Stappers and Pasman (1999) focused on a retrieval results set from a search engine and demonstrated the advantage of visualizing the retrieved documents in the MDS visual space over a linear search-results presentation. Thelwall (2002) examined university relationships through hyperlinks between a pair of university Web sites using the MDS method. Vaughan (2006) analyzed linguistic and cultural differences using Web co-links on related Web sites in the MDS visual environment.

MDS has many advantages. It is relatively free of any data distributional assumptions. The applied dataset does not necessarily follow any type of statistical distribution such as a normal distribution. This characteristic widens its applications and domain. Some data-analysis methods usually require that data be at an ordinal, interval, or ratio level. Basically, the MDS method is tolerant of any of these data levels. The MDS method is especially appropriate for visualization analysis of a small number of objects.

The analysis of query logs lends itself quite readily to information-visualization techniques such as MDS. To demonstrate the applicability of MDS for this purpose, we propose using MDS techniques for the exploratory analysis of query logs from a public search engine on a popular topic area. One area that has not been investigated to date in detail is that of sports-related queries. Sports are not only a physical activity but also are relevant to health, medical treatment, entertainment, culture, education, community, politics, technology, religion, and consumerism (Borish, 1992). Sports are even commoditized as popular culture (Free & Hughson, 2006). Sports play an important role in modern life. This trend was reinforced with the advent of information technology such as the Internet, leading to the tremendous increase in sport's popularity. More and more people search sports-related information via the Internet. The popularity of the Internet has shaped distinct information-seeking behavior; however, this information-seeking behavior has not been fully studied or understood.

The objectives of this research, therefore, are to (a) better understand Internet sports information-seeking behavior in terms of query keywords, (b) shed light on how different sport and related query terms are associated, (c) identify popular sports appearing in queries, and (d) reveal associative subjects and topics of a sport.

The findings of this study could benefit sports searchers who can utilize a related keyword cluster to formulate, refine, or optimize their queries; sports-related businesspeople who may find and explore new business opportunities by identifying relevant sports products and merchandise; and Webmasters and designers of sports-related portals for optimizing their Web information contents and information organization to attract more online users to use their Web sites.

Methods and Analytical Techniques

Query Processing

When users' information needs are converted to queries and these queries are submitted to a search engine, all search keywords in the queries are recorded. The fact that keywords appear in the same query indicates that the searcher has identified some association among the keywords. The keyword semantic associations among all search queries in a transaction log can be revealed if their relations can be identified, organized, and presented in a meaningful way. In fact, the co-occurrence of two associated search keywords in a log file is an indication to some degree of the semantic strength between the keywords. If the co-occurrence between any two keywords can be converted into a form of similarity measure, then proximity relationships among the search keywords can be defined. Exploratory statistical techniques can provide these measures. In particular, forms of clustering analysis applied to search keywords can be conducted using traditional and visualization approaches, much in the same way they have been applied for author co-citation analysis. Traditional and visualization approaches, which are usually complementary to each other in terms of data-analysis accuracy, data presentation, and data interpretation, have their own strengths and weaknesses; therefore, applying both to search keyword clustering analysis would not only confirm results of both approaches but also provide results from quite different perspectives.

The query data used in this study come from the publicly available Excite search engine dataset from December 1999, which consists of 622,785 submitted queries, after requests for additional pages of results for the same query were removed. Although this dataset is not recent, it represents one of the few such logs that have been made readily available to researchers for investigation. The dataset includes only submissions where an identifier, stored as a cookie on a user machine, was available. Queries submitted from machines on which the browser cookie facility was disabled were not included. The log data includes numeric identifiers for machines, the time when each query was submitted, and the full queries entered by searchers. One advantage of using a general search engine data log is that it covers all topics, from which specific topic areas may be mined.

Tokenizing

After all queries were cleaned and identified from the Web log file, individual keywords were parsed in MS Access

DOI: 10.1002/asi

based on standard nontextual delimiters. URLs and e-mail addresses were treated as singular keywords. Phrases such as *table tennis* and *cross country ski* were treated as two and three keywords, respectively, rather than one composite keyword. A search keyword master file was established, where all search keywords (including both content-bearing words and stop words) and their raw frequencies in the query log file were tallied, resulting in 171,731 keywords. All keywords in the search keyword master file were ranked in descending frequency order for further screening. From this list, the investigators identified meaningful sports-related focus keywords for the study.

Identifying Focus Keywords

A focus keyword is defined as a keyword that represents a subtopic of the study topic (e.g., *golf*). A relatively high raw frequency in the investigated query log was necessary to provide a sufficiently large set of co-occurring search keywords for analysis. Keywords in the search keyword master file were screened and examined carefully by the investigators. Eleven sports-related focus keywords whose raw frequencies were $\geq$100 were selected from the search keyword master list.

All keywords that co-occurred with the focus keyword were identified for each of the focus keywords. This step eliminated many unrelated search keywords and isolated the focus keyword and its co-occurring keywords from other keywords in the query log. As a result, it narrowed the query keyword set to a manageable size for the data analysis.

Stop Word Filtering

Some keywords that co-occurred with the focus keyword were not useful; these were treated as stop words and removed from the dataset. The stop words were classified into two categories. The first category included all functional and grammatical words such as a, an, the, with, of, at, under, across, in, at, with, without, and so on. Words in this category bear no semantic information when treated individually and are meaningless for a clustering analysis even though they co-occur with the focus keyword or other keywords. The second category consisted of content-bearing keywords. These keywords were excluded because their levels of co-occurrence were so low that they would make an insignificant contribution to the clustering analysis. A cutoff point, therefore, was set to remove these keywords. All keywords whose corresponding frequencies were smaller than or equal to the cutoff point also were added to the stop word list. Note that the selection of the cutoff point has an indirect impact on the quality of the MDS analysis results. The greater the cutoff value for the focus keyword co-occurrence, the better the corresponding MDS analysis results were in terms of the stress value (discussed later).

Data Filtering

To produce a plausible visual-clustering analysis, a second important cutoff point was introduced to further eliminate useless keywords for a clustering analysis. The total query keyword co-occurrence of a keyword was defined as the number of other keywords with which the keyword in question co-occurred across the complete set of keywords. For instance, if keyword *A* appears in three queries: (*A*, *B*, *C*, *D*), (*A*, *B*, *E*), and (*A*, *C*, *F*), where *A*, *B*, *C*, *D*, *E*, and *F* represent six different keywords, the total query keyword co-occurrences for keyword A is five because it co-occurs with five different keywords. From the keyword co-occurrence analysis perspective, the higher the total query keyword co-occurrences across the query keyword set, the better the results. This is because the identified keyword co-occurring with more keywords can produce a richer comparison. Therefore, the keywords whose total co-occurrences across the query keyword dataset were smaller than a predetermined threshold were excluded from the co-occurrence analysis. A cutoff point was set for this purpose, and all keywords whose corresponding total co-occurrences across the query keyword set were smaller than or equal to the cutoff point also were discarded. Consequently, they were not considered in the later keyword co-occurrence analysis. Clearly, the higher the cutoff point, the lower the number of keywords that may be included in the later keyword co-occurrence analysis, and vice versa. However, each selected sports-related keyword may vary in statistical characteristics. A high cutoff point of total query keyword co-occurrences for a keyword with a low raw keyword frequency in the query keyword master file would lead to fewer displayed keywords in the visual space. Very few displayed keywords may result in a less meaningful display in the visual space. This suggests that different cutoff points should be adopted for different selected focus keywords to ensure enough keywords are included in the later visual keyword co-occurrence analysis. Another factor that affected the selection of the cutoff points was SPSS, the statistical software package used in this study (Version 16.0, 2007). SPSS limits the number of variables to 100, which set the upper limit for the number of keywords that could be included for each analysis. It also kept the visual analysis more manageable.

Matrix Generation and MDS

For those keywords that remained after the initial processing decisions, a co-occurrence matrix was created containing the pairwise frequencies of occurrence between the remaining keywords. This co-occurrence matrix served as the input for the cluster and MDS analyses. Note that the selection of a cutoff point for the total query keyword co-occurrences of each keyword has a direct impact on the quality of the MDS outcome measured by the stress value, which serves as an indication of the goodness of fit (Further details about the stress value are discussed later.) Based on a pilot study that examined the influence of different cutoff values, the investigators noted that as the cutoff value increased, the corresponding stress value decreased. In other words, the keywords that co-occurred with a large number of other keywords tended to produce a better goodness of fit for the analysis. For example, when the cutoff points for total query

DOI: 10.1002/asi

keyword co-occurrences for the focus keywords were set to 4, 5, and 6, the corresponding stress values were 0.09928, 0.09670, and 0.06520, respectively. However, as the value of the cutoff point of the total query keyword co-occurrences of a keyword increased, the number of co-occurring keywords included for visual analysis decreased dramatically. When the cutoff points for the total query keyword co-occurrence of the keyword *Sports* were set to 3, 4, 5, and 6, the numbers of remaining co-occurring keywords for a visual analysis were 59, 42, 27, and 14, respectively. In undertaking the analysis, the investigators kept a reasonably large number of displayed keywords as long as the group of keywords could generate an acceptable stress value.

In summary, the final cutoff point selection for total query keyword co-occurrences for a given keyword was affected by the raw frequency of the focus keyword, the statistical software package limitations, the visual keyword co-occurrence analysis effect, the number of displayed keywords, and the quality of the MDS analysis.

Matrix Normalization and Conversion

A keyword co-occurrence matrix as the raw input cannot be employed directly for the visualization and cluster analysis for a variety of reasons. A normalization process must be applied to reduce the impact of the scale due to different frequencies of the query keywords. A keyword that occurs 100 times should not unduly affect the relationship outcome when compared with a keyword that occurs only 10 times, particularly if their co-occurrences are proportionately distributed. Further, the normalized keyword co-occurrence matrix has to be converted into a similarity matrix. The similarity matrix differs from the keyword co-occurrence matrix, although both matrices are the same in size and structure ($N \times N$ matrices), and the former is derived from the latter.

The similarity between two query keywords is defined as:

$$S(w_i, w_j) = \frac{F_{ij}}{MIN(F_i, F_j)} \tag{1}$$

In Equation 1, w_i and w_j are two keywords in the query log file. $S(w_i, w_j)$ is the similarity between the two keywords w_i and w_j. Legitimate values of $S(w_i, w_j)$ range from 0 and 1. F_{ij} is the number of the queries in which keywords w_i and w_j co-occur. F_i is the number of queries in which keyword w_i occurs; F_j is the number of queries in which keyword w_j occurs. $MIN(x, y)$ is the smaller of the two values x and y.

The similarity matrix is converted into a dissimilarity matrix as the input for the multidimensional scaling visual analysis by the following equation:

$$DS(w_i, w_j) = 1 - S(w_i, w_j) \tag{2}$$

It is apparent that the legitimate values of $DS(w_i, w_j)$ also fall between 0 and 1 after conversion. After the dissimilarity matrix is established as input for the analysis, keywords can be projected onto the visual space for visual analysis.

Stress Value and Loss Function

Because the MDS stress value indicates the quality of an MDS visual display result, a loss function, Kruskal Stress or goodness-of-fit value, was used to calculate the stress value for this study. A loss function defined by the least-squares method is a normalized sum of projection errors over all pairs of objects (i.e., keywords). According to Kruskal (1964a), the general guidelines for stress values outline 20% as poor, 10% as fair, 5% as good, 2.5% as excellent, and 0% as an exact match. Therefore, a smaller stress value indicates a better MDS result. Following this principle, a stress value below about 10% (or 0.1) is considered to be acceptable and plausible for this study. In addition to the stress value, SPSS produces the squared correlation index RSQ (R^2) for the MDS analysis, which is defined as the squared correlation of the input distances with object distances in the MDS visual space. The greater the R^2 value in the visual analysis result, the better the MDS result. When R^2 is greater than or equal to 0.90, the results are usually considered acceptable and plausible. A good and acceptable stress value (or squared correlation index) ensures robust and sound results in the MDS visual space.

Several factors can affect a stress value of an MDS analysis. The Minkowski metric (Korfhage, 1997) can be used to calculate the distance between two objects in the visual space. This measure, in turn, can be used to compute the goodness of fit of the involved objects/keywords in the visual space. The selection of the Minkowski metric from a series of Minkowski families can affect the stress value of the MDS analysis. Given $x_i = (a_1, a_2, \ldots, a_n)$ and $x_j = (b_1, b_2, \ldots, b_n)$, representing two objects in an n-dimensional space, the distance between $x_i = (a_1, a_2, \ldots, a_n)$ and $x_j = (b_1, b_2, \ldots, b_n)$ is defined as:

$$d_{ij} = \left(\sum_{r=1}^{n} (a_r - b_r)^k \right)^{\frac{1}{k}}, \quad k = 1, \ldots, \infty \tag{3}$$

where k is the Minkowski parameter and n is the dimensionality of the visual space.

Based on our pilot study, we found that a lower Minkowski parameter k in Equation 3 tended to generate a better stress value. In the present study, the Minkowski parameter k was set to 1, 2, or 3.

Another factor that affects the final stress value of the MDS analysis is the dimensionality of the MDS visual space. Our pilot study showed that the higher the dimensionality of the visual space, the better the final stress value. However, for a visual-clustering analysis, the MDS results are no longer visible if the dimensionality is higher than 3. For this reason, the eligible dimensionalities of the visual space in this study were 2 and 3.

Note that the MDS feature in SPSS allows users to rotate a three-dimensional visual configuration in the visual space. Consequently, it enables users to observe the displayed clusters from any angle in the visual space. It is extremely

DOI: 10.1002/asi

important to distinguish adjacent clusters, which may overlap in the visual space if they are viewed from a fixed angle.

One of the most important characteristics of the visual-clustering-analysis method is its visual display of the investigated objects. The MDS visual space in this study is three-dimensional while a traditional clustering method results in a one-dimensional space. Two or three-dimensional presentations can illustrate richer information than can a one dimensional presentation. There may be close associations among the objects, which may result in close proximities of the objects in the visual presentation. A good clustering-analysis method not only clearly identifies hidden clusters in a given dataset but also demonstrates sophisticated relationships among the identified clusters and connections among the objects in a cluster. Due to the inherent weakness of traditional clustering-analysis methods, using clustering analysis cannot achieve the latter requirement. Because of its three-dimensional display capability, the MDS visual-clustering method can clearly and intuitively reveal detailed, sophisticated, and multiple relationships among the identified clusters/objects in the visual space. In addition, the visual-clustering method enables people to observe the emerging pattern of clusters and to view the shape of a cluster and the distribution of the objects within that cluster. The MDS visual-clustering-analysis method provides an effective quality-control mechanism for cluster analysis. The stress value of an MDS analysis result, in conjunction with the squared correlation index, *RSQ* (R^2), can indicate the quality of the clustering-analysis result. A control on both the stress value and the squared correlation index ensures a high-quality result for the visual-clustering analysis.

A weakness of visual-clustering analysis is its limited ability to effectively display thousands of objects in its space. Overloaded objects in a limited display area make the cluster determination difficult, if not impossible. For this study, the average number of investigated keywords was less than 100, which is a manageable size for the visual-clustering analysis. The determination of a cluster in the visual environment is flexible and intuitive, but this ability can be a "double-edged sword." When objects are projected together onto the visual space and yield clusters that do not have clear-cut boundaries, it can be difficult to define and identify the clusters in the visual space. However, one of the salient advantages of a traditional clustering-analysis method is that it can clearly define and show the cluster membership for a given dataset regardless of the hidden cluster distribution. Therefore, combining the visual clustering-analysis method with a traditional clustering-analysis method can provide complementary results. Highlighting the clusters defined by the traditional clustering-analysis method in the contexts of the visual-clustering method effectively solves the problem of blurred boundaries.

Although the visual clustering-analysis method has advantages over traditional hierarchical-clustering methods (e.g., the average linkage clustering method and Ward's method), traditional hierarchical-clustering methods also were applied to the datasets in this study for the purpose of comparison. Using a hierarchical-clustering method, the clustered objects are not partitioned in a single step. Instead, a series of partitions takes place within clusters to create multiple-level nested clusters. In this study, an agglomerative hierarchical-clustering algorithm was used to identify clusters from a produced dendrogram. In fact, the number of clusters can be controlled by adjusting the distance between clusters when they are joined in the dendrogram. The greater the selected distance, the fewer the identified clusters. The average linkage clustering method was used for the focus keywords, and the resultant clusters were marked and highlighted in the corresponding MDS visual contexts. The distance between two objects was measured and calculated by the Minkowski metric, with the Minkowski parameter k set to 2 or 3. The hierarchical clusters for each focus keyword and their related keywords were identified in the corresponding dendrogram.

Another traditional clustering method, Ward's clustering method, also was applied to the investigated focus keywords to confirm whether the clustering-analysis results of the two methods were consistent. To maintain good readability for the visual displays, we did not include the clusters created by the Ward's clustering method in the visual displays as we did for the average linkage clustering method.

In this study, two traditional clustering methods were employed as two comparison baselines to corroborate the clustering-analysis results from the visual MDS method. In Ward's method, clusters are generated in such a way that the squared distance to the center mean is minimized; in the average linkage method, objects are clustered based on the average distance between all pairs of objects. Both the average linkage and Ward's methods are different members of the hierarchical agglomerative clustering-analysis family. A hierarchical method allows us to control the number of generated clusters, which facilitates the comparison between the results of the visual MDS method and those of a traditional clustering method.

By superimposing the results of the two traditional clustering methods and the MDS visual-clustering method, the outcomes of the two approaches may be compared. Note that because different methods have different grouping criteria, the results of the each method might be slightly different. Any such differences are discussed for each focus keyword.

Experimental Study

To examine whether the related terms in a cluster yielded from the users' queries in this study help users enhance search effectiveness, the authors conducted a small-scale experimental study. Search tasks for the experimental study were developed based on the investigated focus keywords (*golf*, *sports*, *ski*, *football*, *wrestling*, *hockey*, *soccer*, *baseball*, *tennis*, *boxing*, and *bowling*). Information-need descriptions associated with each of the search tasks were provided to searchers/participants. Each participant was asked to conduct two searches in Google. In one search, the participant completed his or her search without any assistance; in the other search, the participant completed his or her search with

DOI: 10.1002/asi

TABLE 1. Summary of focus keywords and cutoff points.

Focus keyword	Raw frequency in the log file	Stop word list size	*CPCFKAK*	*CPTQKC*
Golf	766	680	4	4
Football	601	532	3	4
Sports	744	756	3	3
Ski	765	536	3	4
Wrestling	495	413	3	3
Hockey	313	326	3	3
Baseball	278	309	3	3
Soccer	285	289	3	3
Bowling	126	96	2	2
Tennis	126	125	2	2
Boxing	104	105	2	2

TABLE 2. Summary of the focus keyword investigation.

Focus keyword	Stress value	RSQ (R^2)	Keyword group size	Minkowski parameter
Golf	0.0617	0.99514	37	2
Football	0.08972	0.99005	82	2
Sports	0.07272	0.99332	59	2
Ski	0.05808	0.99608	67	2
Wrestling	0.07984	0.99199	60	2
Hockey	0.08424	0.99041	39	3
Baseball	0.07797	0.9852	22	2
Soccer	0.09499	0.9833	29	2
Bowling	0.09461	0.98248	23	1
Tennis	0.09014	0.98868	24	1
Boxing	0.06840	0.99317	23	2

the assistance of a search-term list generated from relevant queries in the transaction log. A *t* test was employed to check whether the difference between the two groups of search performance was statistically significant.

The participants in this study were 2 graduate students in information studies and in architecture and urban planning at the University of Wisconsin, Milwaukee. They were familiar with basic information-search skills.

Results

Focus Keyword Dataset Characteristics and Outcome Summary

Based on the research method described in the previous section, 11 focus keywords were identified and selected from the query log file for study. Again, the keywords were *golf*, *football*, *sports*, *ski*, *hockey*, *wrestling*, *baseball*, *soccer*, *bowling*, *tennis*, and *boxing*. *Basketball* was initially considered as a focus keyword, but satisfactory MDS outcomes could not be attained so it was dropped from the analysis. The corresponding raw frequencies of these focus keywords are listed in Table 1. The maximum and minimum raw frequencies for this group of focus keywords were 766 for *golf* and 104 for *boxing*, respectively. The average raw frequency was 418.45. The maximum and minimum stop word list sizes were 680 for *golf* and 96 for *bowling*, respectively, with an average stop word total of 378.82 keywords. The stop word list size of a focus keyword corresponded roughly to its raw frequency in the query log. In other words, a focus keyword with a higher raw frequency in the query log file would correspond roughly to a larger stop word list.

To define a group of related keywords for a focus keyword, the two cutoff points outlined earlier were determined. The cutoff points for the co-occurrence of a focus keyword and an associated keyword (*CPCFKAK*) and the cutoff point for the total query keyword co-occurrence (*CPTQKC*) of a keyword across all query keyword sets are listed in Table 1. *CPCFKAK* and *CPTQKC* values range from 2 to 4. Note that a focus keyword with a high raw frequency tended to have both a higher *CPCFKAK* cutoff point and a higher *CPTQKC* cutoff point.

TABLE 3. Summary of the hierarchical clustering analysis.

Focus keyword	DLBC	MP	NIC
Golf	8	2	5
Football	6	2	4
Sports	5	2	4
Ski	12	4	3
Wrestling	11	3	4
Hockey	8	2	5
Baseball	11	2	3
Soccer	8	2	4
Bowling	11	2	3
Tennis	11	2	3
Boxing	12	2	3

DLBC = distance length between clusters in a dendrogram; MP = Minkowski parameter; NIC = the number of identified clusters in a focus keyword and its related keywords.

Each focus keyword ultimately defines a group of associated keywords extracted from the query log. The sizes of the associated keyword groups ranged from 23 (*boxing*) to 82 (*football*). The average size was 42.27. The Minkowski parameters for the MDS visual-clustering analysis were set to 1, 2, or 3 in this study. The Minkowski parameter was set to 2 for eight of the focus keywords, 1 for two focus keywords, and 3 for one focus keyword. No MDS analysis result stress value was greater than 0.10000. The greatest stress value was 0.09499 (*soccer*), and the smallest stress value was 0.05808 (*ski*). The average stress value for this group of the focus keywords was 0.07931. All RSQ (R^2) values in this investigation were greater than 0.9. The highest and the lowest RSQ values were 0.99608 (*ski*) and 0.98248 (*bowling*), respectively (Table 2). The average RSQ value for these focus keywords was 0.989983636. Both the stress values and RSQ values for all the selected focus keywords demonstrate that the experiment achieved satisfactory and sound results.

The final results of the hierarchical clustering analysis for the 11 focus keywords appear in Table 3. A detailed presentation of the results follows. Due to space limitations in the figures themselves, the displayed query keywords are represented by abbreviations (*Vxx*), with corresponding

DOI: 10.1002/asi

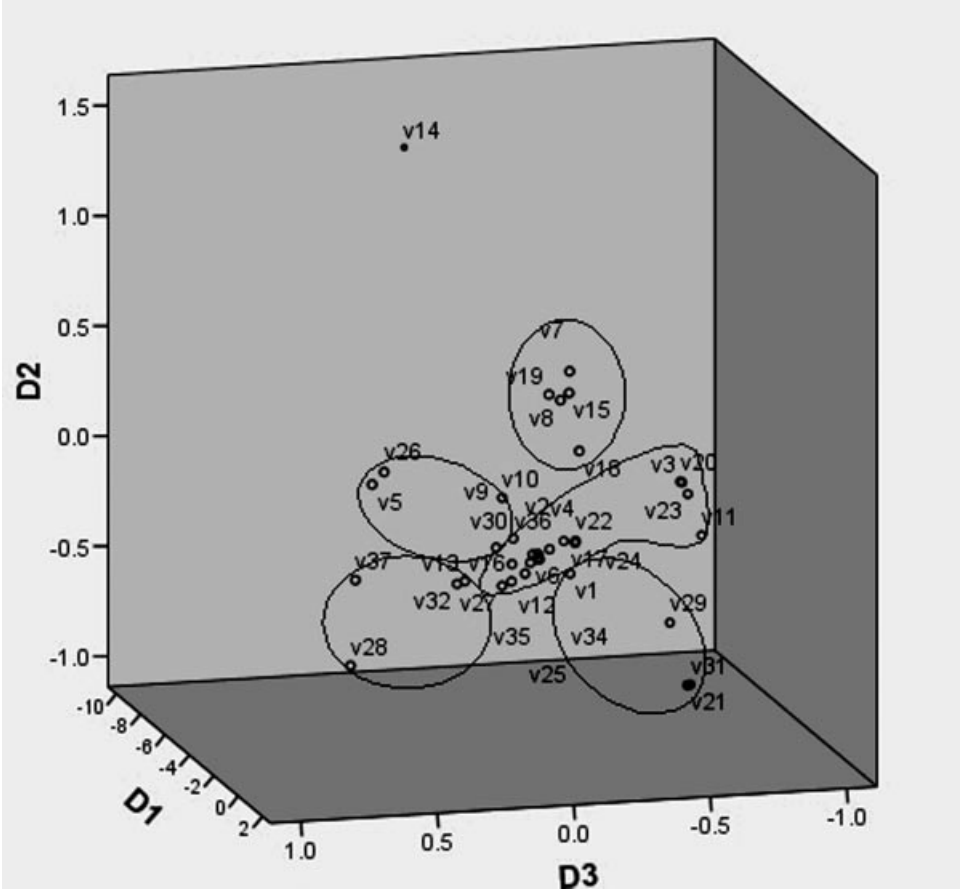

FIG. 1. Visual display for focus keyword *golf*.

keywords for each abbreviation appearing below the figure. The assigned membership of each keyword in the generated clusters is identified with closed-line drawings.

Golf

Five clusters (C1–C5) from the average linkage clustering method were identified (Figure 1), with the following members:

C1: *open*, *U.S.*, *tournament*, *2000*
C2: *Tiger*, *Woods*, *game*, *ultimate*
C3: *Carolina*, *club*, *course*, *south*
C4: *Charlotte*, *Charweb*, *golfing*, *LFING*, *Mecklenburg*
C5: *Arizona*, *beach*, *California*, *carts*, *Diego*, *equipment*, *greens*, *irons*, *Myrtle*, *packages*, *San*, *school*, *senior*, *state*, *travel*, *used*, *vacations*, *Washington*, *women*

The stress and RSQ values indicate a very good fit. Not surprisingly, *Tiger* and *Woods*; *U.S.*, *tournament*, and *open*; *travel*, *vacation*, *package*, and *beach* are grouped together, respectively. Because the focus keyword *golf* (V14) is supposed to relate to all included query keywords, it is not grouped into any of the clusters in the visual space. C5 is the largest cluster, which is located at the center. Some states such as *Washington*, *California*, and *Arizona* appeared in this cluster. Within C5, *beach* (V3), *Myrtle* (V20), and *vacations* (V34) formed a subcluster because Myrtle Beach is a popular vacation place. Another meaningful subcluster within C5 consists of *San* and *Diego*.

For the focus keyword *golf*, the resultant clusters from Ward's clustering method are listed as follows.

C1′: *open*, *U.S.*, *tournament*
C2′: *Tiger*, *Woods*, *game*, *ultimate*
C3′: *Carolina*, *club*, *course*, *greens*, *south*
C4′: *Charlotte*, *Charweb*, *golfing*, *LFING*, *Mecklenburg*
C5′: *2000*, *Arizona*, *beach*, *California*, *carts*, *Diego*, *equipment*, *irons*, *Myrtle*, *packages*, *San*, *school*, *senior*, *state*, *travel*, *used*, *vacations*, *Washington*, *women*

Note that the two traditional clustering methods yielded almost the same results for all but two keywords. The keyword *2000* in C1 from the average linkage clustering method is in C5 from Ward's clustering method, and the keyword *greens* in C5 from the average linkage clustering method is in C3′ from Ward's method.

Football

The focus keyword *football* (V27) generated 82 related keywords for the analysis (Figure 2), and produced acceptable

DOI: 10.1002/asi

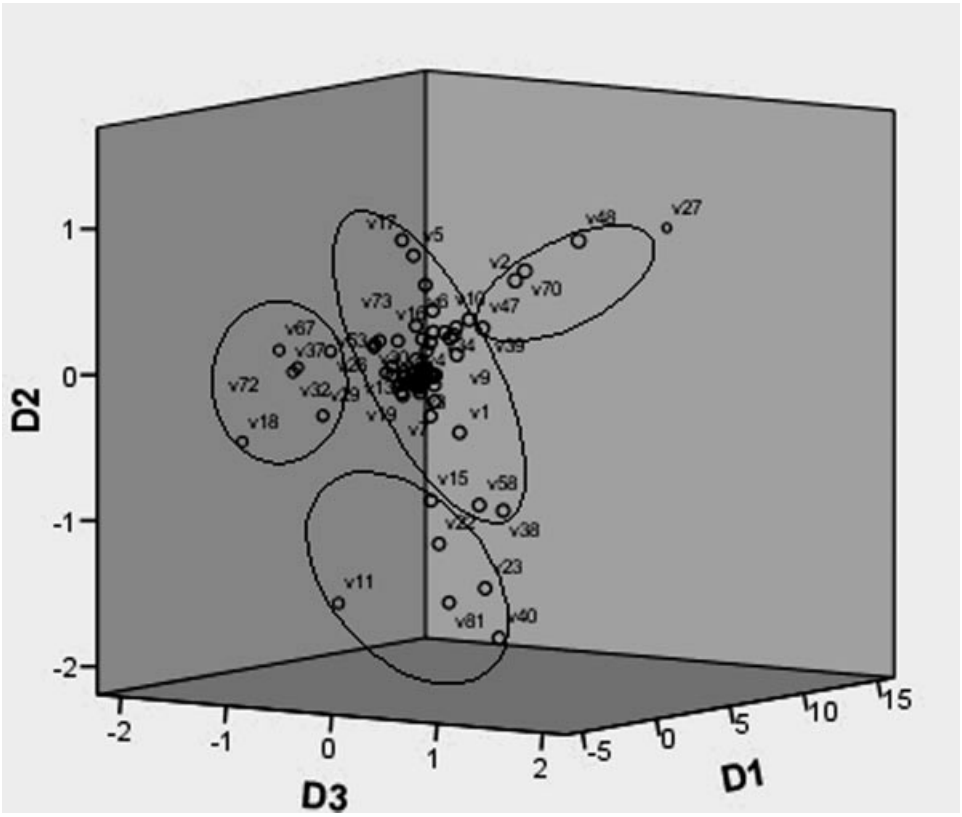

FIG. 2. Visual display for focus keyword *football*.

stress and RSQ values. Four clusters from the average linkage clustering method emerged:

C1: *championship*, *computer*, *Eidos*, *manager*, *winners*, *FA*
C2: *AFC*, *NFC*, *NFL*, *Sportsbook*
C3: *high*, *defeats*, *games*, *state*, *school*, *Kentucky*
C4: *2000*, *American*, *apparel*, *badger*, *basketball*, *bay*, *betting*, *bowl*, *cards*, *club*, *coaching*, *college*, *Dame*, *Dayne*, *demo*, *division*, *Duke*, *fantasy*, *flag*, *Florida*, *free*, *Gators*, *helmets*, *history*, *hockey*, *jacket*, *jerseys*, *league*, *Letterman*, *merchandise*, *Michigan*, *Minnesota*, *Monday*, *national*, *NCAA*, *night*, *Notre*, *Ohio*, *Oklahoma*, *Penn*, *picks*, *pictures*, *players*, *playoffs*, *premier*, *pro*, *rankings*, *records*, *recruiting*, *Riddell*, *rosters*, *rules*, *schedule*, *scores*, *sports*, *standings*, *team*, *Texas*, *trading*, *united*, *university*, *UW*, *week*, *wholesale*, *Wisconsin*

C4 is the largest cluster, and most related keywords fall into this group, and not unexpectedly, it is situated at the center of the visual space. It is clear that professional associations (*AFC*, *NFC*, and *NFL*) are grouped together in C2, along with an online sports-betting Web site, *Sportsbook*. In C4, there are several meaningful subclusters; for example, *Letterman*, *jacket*, and *wholesale* produced a subgroup; *cards*, *trading*, *basketball*, and *hockey* formed another subgroup; and *UW*, *badger*, and *Dayne* generated a subgroup. Keyword pairs *Florida* and *gators*, *NCAA* and *rules*, and *Notre* and *Dame* formed subgroups, respectively.

The resultant clusters from Ward's clustering method are the same as those from the average linkage clustering method.

Sports

This focus keyword *sports* (V46) is more general than are the other 10 selected focus keywords. The analysis generated 59 related keywords (Figure 3). The final test stress value and RSQ value were quite good. The four clusters from the average linkage clustering method were clearly identified:

C1: *Aggroup*, *SFX*, *management*, *group*, *onlinesports*
C2: *football*, *history*, *refereeing*
C3: *illustrated*, *Swimsuit*, *Issue*, *edition*
C4: *American*, *art*, *athletes*, *Canada*, *car*, *century*, *city*, *classic*, *Cleveland*, *clip*, *clubs*, *Columbia*, *eastern*, *ESPN*, *events*, *excite*, *fox*, *free*, *injury*, *Iowa*, *jobs*, *magazines*, *Minnesota*, *Missouri*, *motor*, *news*, *players*, *pro*, *radio*, *recruiting*, *royal*, *san*, *school*, *scores*, *shop*, *sports*, *stars*, *state*, *store*, *swim*, *teams*, *ticket*, *training*, *university*, *water*, *women*, *York*

The keywords *edition*, *issue*, and *illustrated* are grouped together to reflect the popular sports magazine "Sports Illustrated." The keywords *shop* and *store* indicate possible sports-related online shopping activities.

Within the largest cluster, C4, *injury* and *training* were tied together; *athletes*, *recruiting*, and *jobs* were closely related; and *classic* and *car* were paired.

DOI: 10.1002/asi

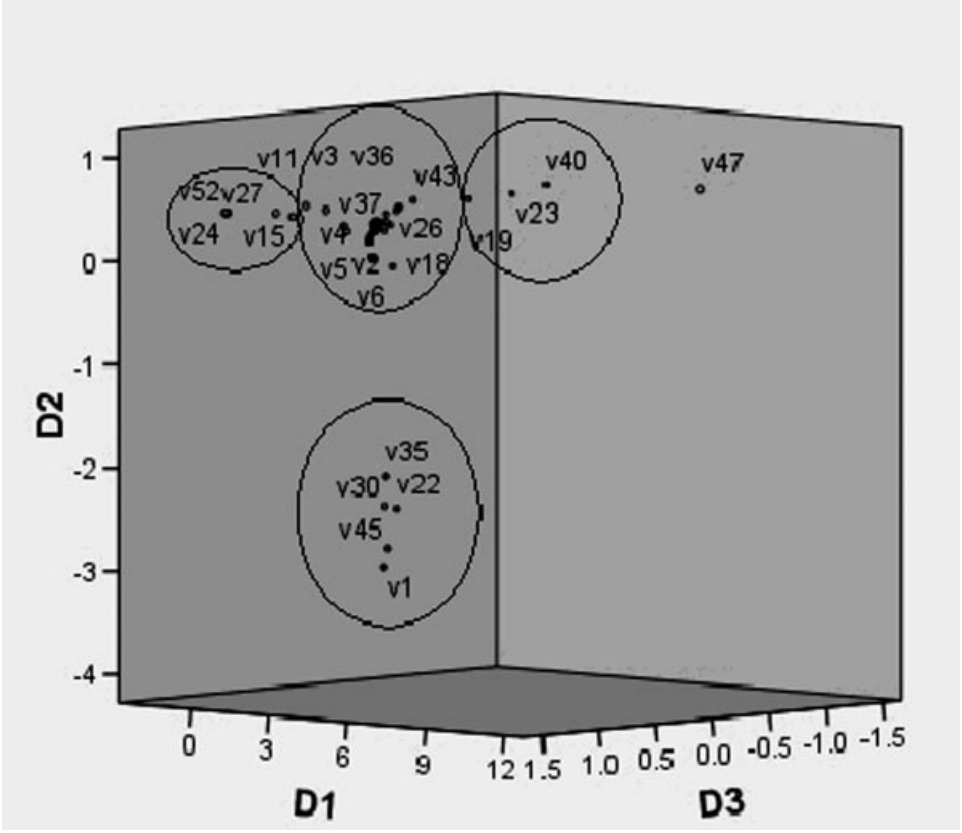

FIG. 3. Visual display for focus keyword *sports*.

The resultant clusters from Ward's clustering method are the same as those from the average linkage clustering method.

Ski

Sixty-seven relevant keywords were identified for the focus keyword *ski* (V48). The MDS outcome resulted in very good stress and RSQ values. Three clusters from the average linkage clustering method emerged, represented in Figure 4:

C1: *wolf*, *creek*
C2: *resort*, *west*, *Virginia*, *snowshoe*
C3: *Alberta*, *Alpine*, *area*, *association*, *basin*, *bear*, *boreal*, *bowl*, *California*, *Canada*, *clothes*, *clothing*, *club*, *Colorado*, *conditions*, *country*, *downhill*, *face*, *gear*, *hill*, *hood*, *Kirkwood*, *lake*, *lift*, *lodging*, *Maine*, *mammoth*, *meadows*, *Mexico*, *Michigan*, *Montana*, *mount*, *mountain*, *MT*, *new*, *Norstar*, *north*, *online*, *packages*, *powder*, *rental*, *report*, *rib*, *ridge*, *Sierra*, *ski*, *snow*, *snowboard*, *springs*, *state*, *summit*, *sun*, *Tahoe*, *Taos*, *tickets*, *trails*, *vacations*, *Vail*, *valley*, *weather*, *Whistler*

The outdoor nature of this sport is reflected in the keywords and the generated clusters, such as *valley*, *snow*, *summit*, *mountain*, *sierra*, *sun*, *weather*, *spring*, *hill*, and *downhill*. Also evident was the strong relationship of keywords related to special equipment for skiing, including *gear*, *lift*, *snowboard*, *clothes*, and *clothing*. The first two clusters represent prominent skiing areas that did not share as many keywords, unlike similar keyword groupings that appeared in the largest cluster, C3, in which *California*, *Sierra*, and *summit* were grouped together. In addition, it is no surprise that *North*, *Face*, *clothes*, and *gear* formed a subgroup because "North Face" is a sports clothing brand.

The resultant clusters from Ward's clustering method are as follows.

C1′: *clothes*, *face*, *north*
C2′: *lift*, *norstar*, *online*, *tickets*
C3′: *Alberta*, *Alpine*, *area*, *association*, *basin*, *bear*, *boreal*, *bowl*, *California*, *Canada*, *clothing*, *club*, *Colorado*, *conditions*, *country*, *creek*, *downhill*, *gear*, *hill*, *hood*, *Kirkwood*, *lake*, *lodging*, *Maine*, *mammoth*, *meadows*, *Mexico*, *Michigan*, *Montana*, *mount*, *mountain*, *MT*, *new*, *packages*, *powder*, *rental*, *report*, *resort*, *rib*, *ridge*, *Sierra*, *ski*, *snow*, *snowboard*, *snowshoe*, *springs*, *state*, *summit*, *sun*, *Tahoe*, *taos*, *trails*, *vacations*, *Vail*, *valley*, *Virginia*, *weather*, *west*, *Whistler*, *wolf*

Although each method created three clusters, there were differences between the results. C1′ and C2′ from Ward's clustering method are totally different from C1 and C2 for the average linkage clustering method; however, observe that C1′ and C2′ are two subclusters of C3, and C1 is a subcluster of C3′. In fact, the visual display in Figure 4 also confirms that the keywords were not spread out, and there is no clear distinction between the generated clusters.

DOI: 10.1002/asi

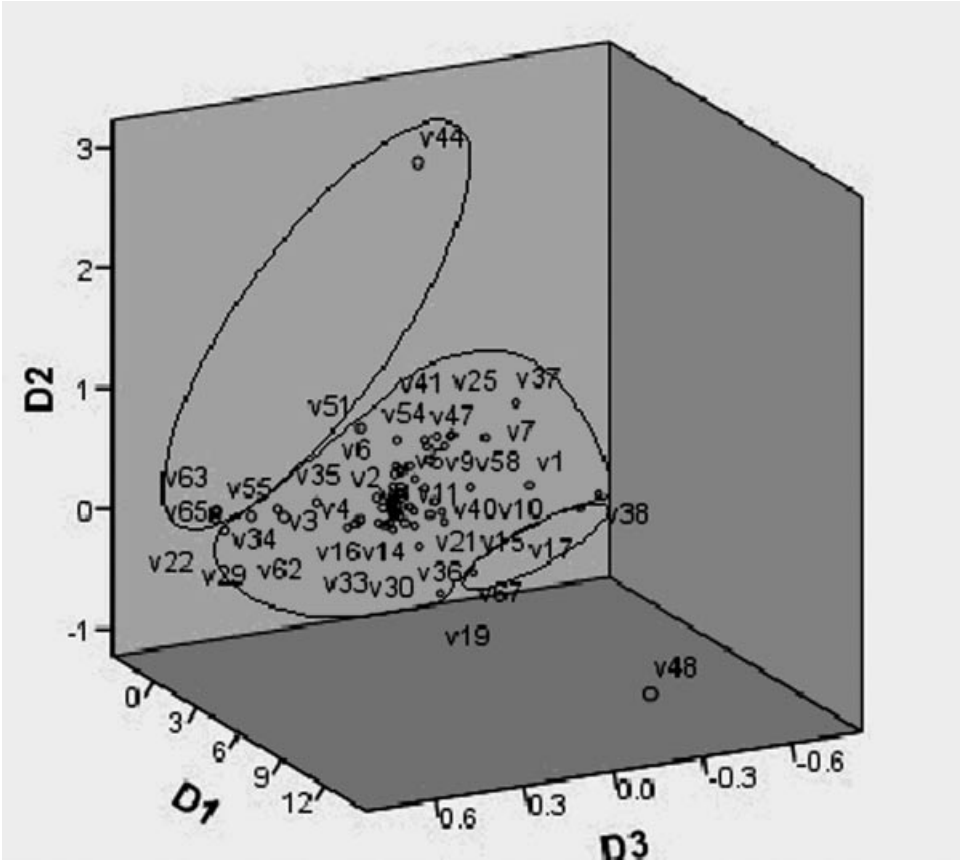

FIG. 4. Visual display for focus keyword *ski*.

Wrestling

Sixty relevant keywords were identified for the focus keyword *wrestling* (V57), and resulted in good stress and RSQ outcomes. Four clusters from the average linkage clustering method were identified (Figure 5):

C1: *screen*, *savers*
C2: *gooey*, *messy*, *mud*
C3: *professional*, *entrance*, *matches*, *music*, *real*, *death*
C4: *Atlantic*, *beast*, *bush*, *championship*, *Chicago*, *east*, *erotic*, *federation*, *female*, *free*, *gay*, *girls*, *high*, *Illinois*, *Iowa*, *ladies*, *links*, *male*, *Missouri*, *mixed*, *naked*, *news*, *nude*, *page*, *Pennsylvania*, *photos*, *pics*, *pictures*, *rankings*, *resource*, *results*, *rumors*, *sable*, *school*, *schools*, *sites*, *sunny*, *train*, *training*, *valets*, *WCW*, *Web*, *weight*, *woman*, *women*, *world*, *wrestling*, *WWF*, *youth*

The appearance of *screen savers* was somewhat unexpected. People were searching for wrestling-themed screen savers on the Internet, but these keywords did not co-occur frequently with any other investigated terms than wrestling. This would explain why the two keywords appear in their own cluster. The same is true for C2. C3 represents other labels associated with wrestling or themes such as *music*.

Within the largest cluster, C4, the two primary professional wrestling associations at the time, *WCW* and *WWF*, were grouped together along with their expanded keywords *federation*, *world*, and *championship*.

The resultant clusters from Ward's method analysis are as follows.

C1′: *professional*, *matches*, *death*
C2′: *Chicago*, *Illinois*, *schools*, *train*, *training*, *weight*
C3′: *high*, *school*, *rankings*, *Pennsylvania*
C4′: *gooey*, *messy*, *mud*
C5′: *Atlantic*, *beast*, *bush*, *championship*, *east*, *entrance*, *erotic*, *federation*, *female*, *free*, *gay*, *girls*, *Iowa*, *ladies*, *links*, *male*, *Missouri*, *mixed*, *music*, *naked*, *news*, *nude*, *page*, *photos*, *pics*, *pictures*, *real*, *resource*, *results*, *rumors*, *sable*, *savers*, *screen*, *sites*, *sunny*, *valets*, *WCW*, *web*, *woman*, *women*, *world*, *wrestling*, *WWF*, *youth*

In this case, the number of the clusters and the corresponding elements from both clustering methods are different. Specifically, C1′, C2′, and C3′ from Ward's clustering method are subclusters of C4 from the average linkage method.

The visual display in Figure 5 also reveals there is no explicit boundary for the produced clusters.

Hockey

For the focus keyword *hockey* (V12), there are 39 associated keywords. The final stress value and RSQ value were

DOI: 10.1002/asi

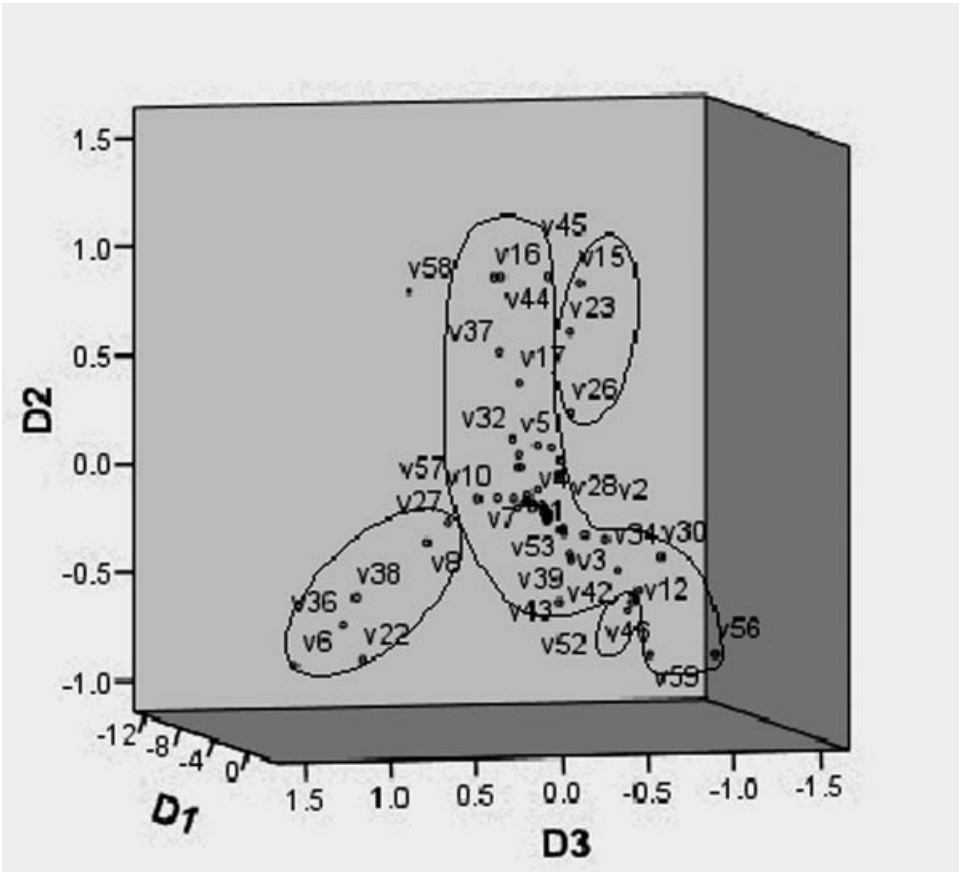

FIG. 5. Visual display for focus keyword *wrestling*.

acceptable. Keywords are grouped into five clusters from the average linkage clustering method (Figure 6):

C1: *island*, *long*, *national*, *league*, *Plymouth*, *MA*, *British*
C2: *north*, *York*
C3: *Toronto*, *hall*, *fame*
C4: *football*, *basketball*, *cards*
C5: *association*, *city*, *equipment*, *game*, *helmets*, *hockey*, *ice*, *international*, *jersey*, *junior*, *Minnesota*, *minor*, *network*, *NHL*, *Ontario*, *pads*, *players*, *roller*, *rules*, *team*, *tournaments*, *wild*, *world*, *youth*

C1 and C2 have geographic themes. Other sports, such as *football* and *basketball*, are found in C4, which contains types of sports memorabilia. The clustering of *Toronto*, *hall*, and *fame* in C3 is not unexpected given the Toronto location of the Hockey Hall of Fame. The largest cluster, C5, represents a mixed bag of keywords, with an equipment-related subgroup consisting of *equipment*, *helmets*, *pads*, and *roller*.

The resultant clusters from Ward's clustering method are listed as follows.

C1′: *island*, *long*, *national*, *league*, *Plymouth*, *MA*, *British*
C2′: *north*, *York*
C3′: *Toronto*, *hall*, *fame*
C4′: *association*, *basketball*, *cards*, *city*, *equipment*, *football*, *game*, *helmets*, *hockey*, *ice*, *international*, *jersey*, *junior*, *Minnesota*, *minor*, *network*, *NHL*, *Ontario*, *pads*, *players*, *roller*, *rules*, *team*, *tournaments*, *wild*, *world*, *youth*

It is apparent that C1, C2, and C3 from the average linkage clustering method are the same as C1′, C2′, and C3′ from Ward's clustering method. The Ward's method analysis did not group C4 and C5 as a cluster, but C4 and C5 are two subclusters of C4′ from Ward's method.

Baseball

Twenty-two related keywords were identified for the focus keyword *baseball* (V1), with good outcomes for the final stress and RSQ values. Three clusters from the average linkage clustering method emerged (Figure 7):

C1: *New*, *York*, *hall*, *fame*
C2: *series*, *world*, *hats*, *history*, *league*, *major*, *MLB*
C3: *baseball*, *caps*, *college*, *fiber*, *little*, *minor*, *Missouri*, *optic*, *pictures*, *player*, *youth*

C1's purpose is similar to that of C3 for *hockey*. Keywords *fiber* and *optic* seem out of place when associated with baseball. But when viewed in the context of fiber-optic baseball caps, their relationship becomes clearer.

The resultant clusters from the Ward's clustering method are as follows.

DOI: 10.1002/asi

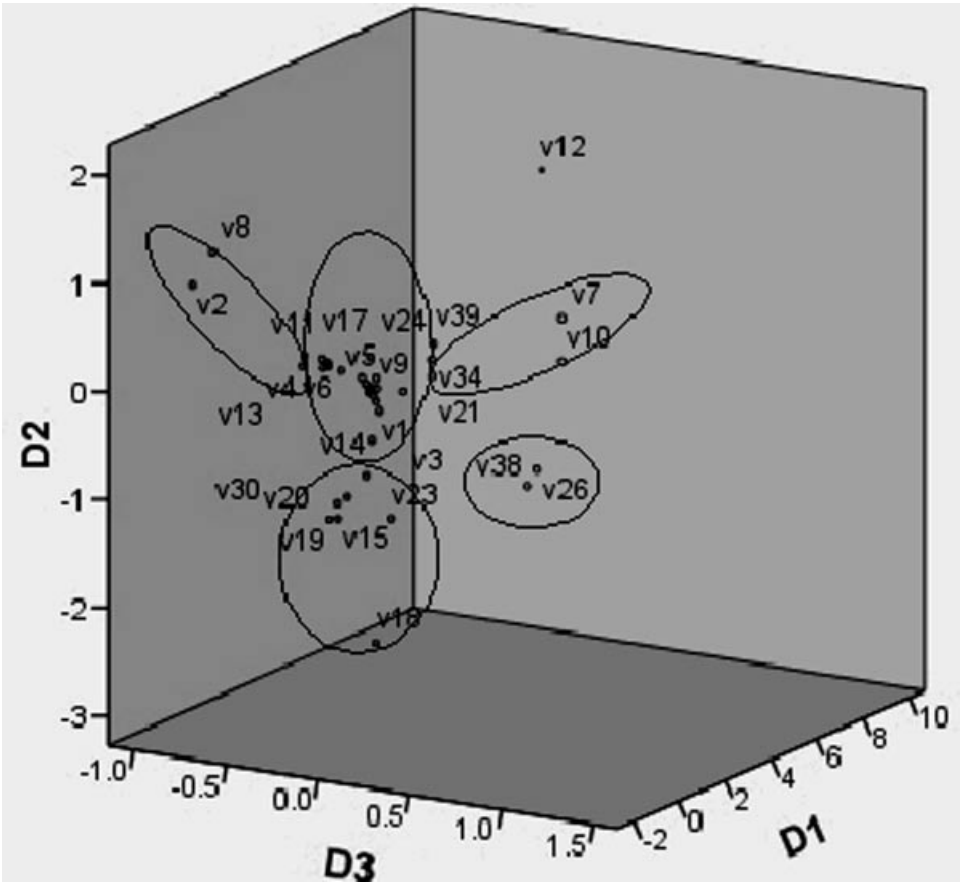

FIG. 6. Visual display for focus keyword *hockey*.

C1′: *York*, *fame*, *new*, *hall*
C2′: *series*, *world*, *history*, *league*, *major*, *MLB*
C3′: *baseball*, *caps*, *college*, *fiber*, *hats*, *little*, *minor*, *Missouri*, *optic*, *pictures*, *player*, *youth*

The two traditional clustering methods produced almost the same resultant clusters except for one keyword. The keyword *hats* is located in C3′ instead of C2′ from Ward's clustering method.

Soccer

For the focus keyword *soccer* (V20), there are 29 relevant keywords. The final stress value and RSQ value were acceptable, but represented the poorest fit of all the focus keywords. Four main groups emerged from the average linkage clustering method (Figure 8):

C1: *high*, *rankings*, *school*
C2: *champion*, *class*, *world*
C3: *Australian*, *Olympic*, *team*
C4: *association*, *calendar*, *California*, *club*, *English*, *girls*, *jersey*, *league*, *Matildas*, *Mexican*, *national*, *nude*, *soccer*, *state*, *teams*, *tournaments*, *US*, *women*, *women's*, *youth*

The first three clusters represent well-defined topics. C4, again, represents a mixed bag of topics.

For the focus keyword *soccer*, the resultant clusters from Ward's clustering method are listed as follows.

C1′: *high*, *school*
C2′: *champion*, *class*, *world*
C3′: *Australian*, *Olympic*, *team*
C4′: *association*, *calendar*, *California*, *club*, *English*, *girls*, *jersey*, *league*, *Matildas*, *Mexican*, *national*, *nude*, *rankings*, *soccer*, *state*, *teams*, *tournaments*, *US*, *women*, *women's*, *youth*

The two traditional clustering methods produced almost the same resultant clusters except for one keyword. The keyword *rankings* is located in C4′ instead of C1′ from Ward's clustering method.

Bowling

Twenty-three relevant keywords were associated with the focus keyword *bowling* (V3). The final stress and RSQ values were acceptable, but among the poorest fitting. Three main clusters from the average linkage clustering method were identified (Figure 9):

C1: *city*, *green*, *Ohio*, *schools*
C2: *Christmas*, *download*, *elf*, *game*, *virus*
C3: *American*, *association*, *balls*, *bowling*, *Brunswick*, *congress*, *hammer*, *KY*, *Santa*, *saver*, *screen*, *shoes*, *state*, *university*

DOI: 10.1002/asi

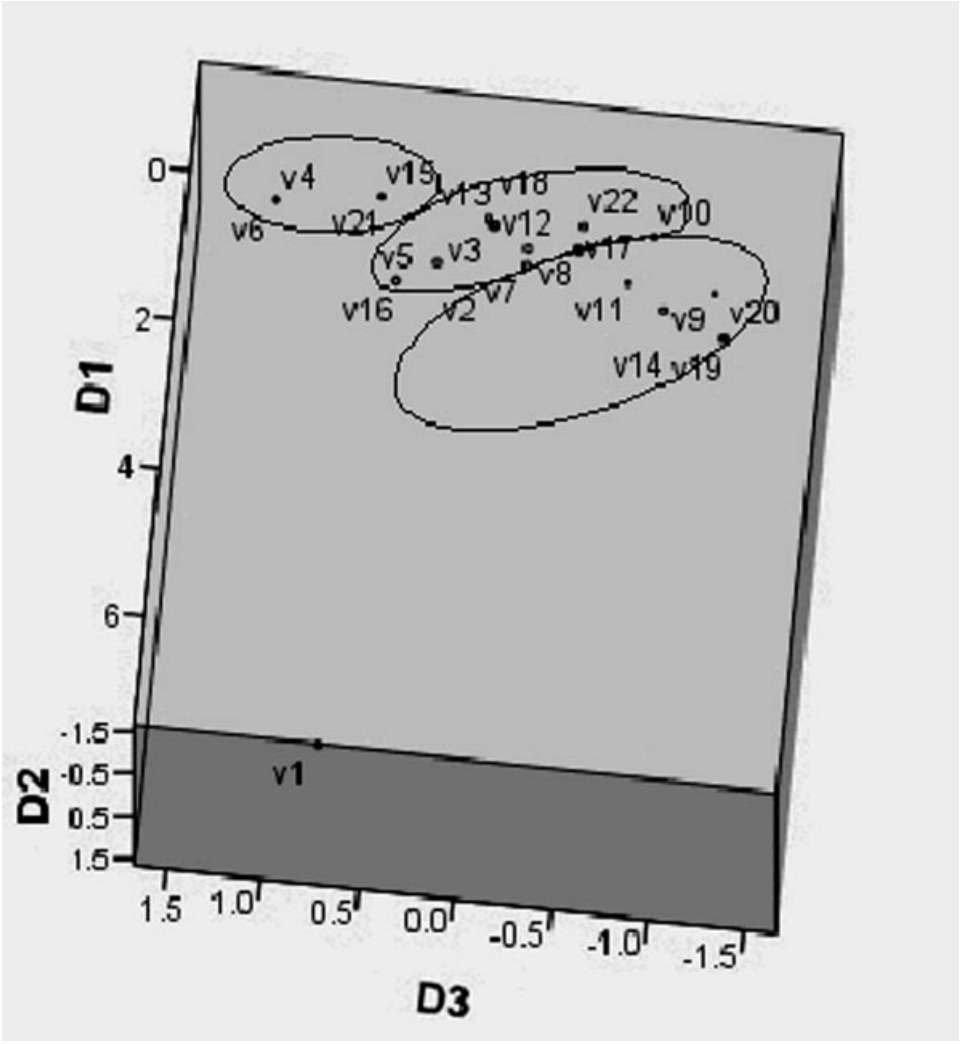

FIG. 7. Visual display for focus keyword *baseball*.

Most keywords in C2 relate to computers. Keywords *Christmas* and *elf* are undoubtedly due to collection time frame for the transaction log, which occurred in December. In C3, *shoes*, *Brunswick*, *balls*, and *hammer* formed a meaningful subcluster, relating to equipment. C1 refers to a non-bowling-related set of keywords; they relate to the city of Bowling Green, Ohio, and schools.

Ward's clustering method is the same as those from the average linkage clustering method.

Tennis

For the focus keyword *tennis* (V19), there are 24 relevant keywords. The final stress and RSQ values were acceptable. The cluster analysis generated four groups from the average linkage clustering method (Figure 10):

C1: *female*, *sexy*, *stars*
C2: *Australian*, *Chris*, *open*, *tournament*
C3: *Adidas*, *association*, *camp*, *coach*, *equipment*, *golf*, *Kournikova*, *Las*, *nude*, *player*, *players*, *table*, *tennis*, *upskirt*, *US*, *Vegas*, *women*

Tennis-related searches appear to deal with the themes of female tennis players, tournaments, and equipment. Observe that C1 appears to be situated inside C3. This is an effect of the two-dimensional representation of the three-dimensional space.

For the focus keyword *tennis*, the resultant clusters from Ward's clustering method are listed as follows.

C1′: *Las*, *Vegas*, *camp*, *golf*
C2′: *female*, *sexy*, *stars*
C3′: *Australian*, *Chris*, *open*, *tournament*
C4′: *Adidas*, *association*, *coach*, *equipment*, *Kournikova*, *nude*, *player*, *players*, *table*, *tennis*, *upskirt*, *US*, *women*

Keywords *Las*, *Vegas*, *camp*, and *golf* were grouped as a cluster in Ward's clustering method while they were part of C3 in the average linkage clustering method.

Boxing

Twenty-three related keywords were identified with the focus keyword *boxing* (V4), with very good final stress and

DOI: 10.1002/asi

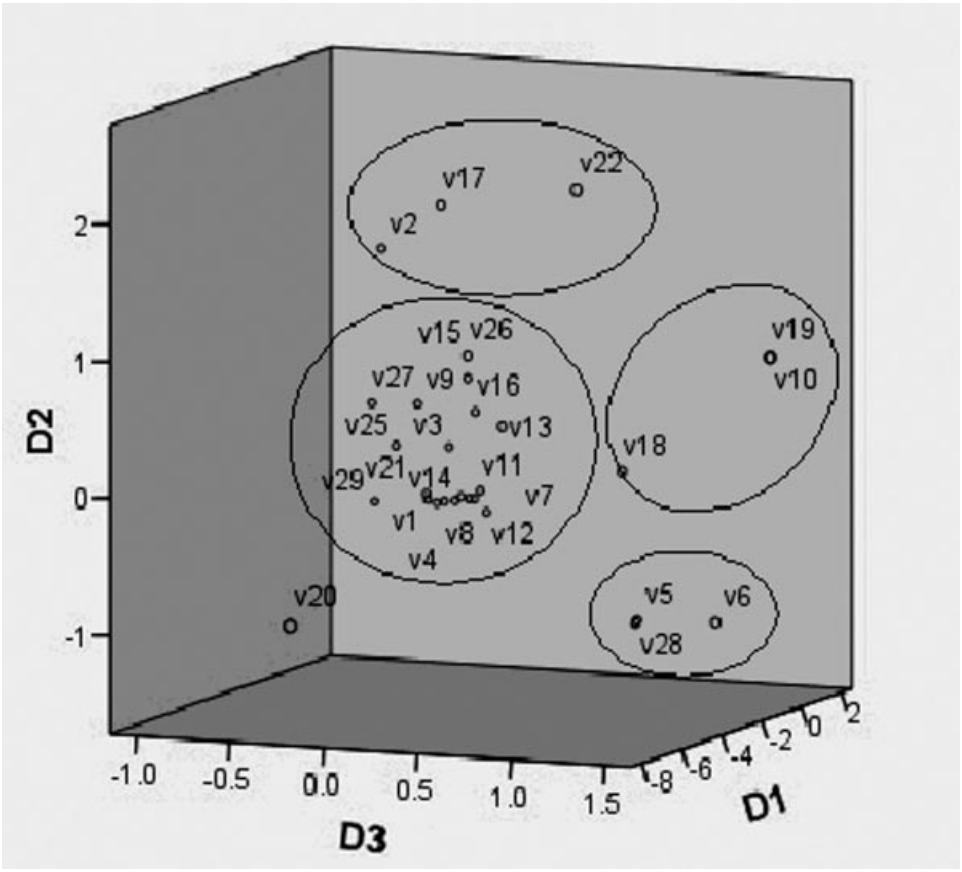

FIG. 8. Visual display for focus keyword *soccer*.

RSQ values. The average linkage clustering method produced three clusters (Figure 11):

C1: *chick*, *topless*, *wife*
C2: *Clay*, *Cooper*, *Henry*
C3: *1930's*, *Baltimore*, *boxers*, *boxing*, *codes*, *day*, *events*, *gun*, *holiday*, *holidays*, *instructions*, *midway*, *naked*, *pictures*, *pre-fight*, *rumble*, *Victorian*

The keywords in C2 are associated with famous boxers while keywords in C1 appear to be related to female boxing.

The resultant clusters from Ward's clustering method are the same as those from the average linkage clustering method.

Categorization of Query Keywords

It is interesting that although the selected sports-related focus keywords vary, they share some commonalities in terms of their related keywords in a query log file. If the related keywords associated with the focus keywords are categorized and characterized, the commonalities can be clearly revealed. After all the focus keywords and their related keywords were compared, common keywords were identified. The related query keywords can be classified into the following seven roughly defined categories: (a) education & training; (b) services; (c) geographic location; (d) professional groups, people, & events; (e) equipment & facilities; (f) computer & information technology (IT); and (g) gender-related. The relationships between the investigated keywords and the categories are summarized in Table 4. All 11 investigated sports-related focus keywords are related to the following categories: geographic location, and professional associations, people, & events, followed by equipment & facilities (eight focus keywords), and education & training (eight focus keywords).

Major themes emerge from this comparison. Among the 11 selected focus keywords, eight include the education & training category. Associated keywords for this category included *school*, *training*, *university*, *camp*, and *instruction*. Service-related query keywords, which appeared in 6 of the 11 focus keyword groups, included *vacations*, *packages*, *merchandise*, *wholesale*, *sportsbook*, *store*, *tickets*, *shop*, *radio*, *rental*, *news*, *valet*, *magazine*, *lodging*, *resort*, and *Christmas*. Geographic location-related keywords, which were found in all of the focus keyword groups, frequently referred to popular locations or events associated with given sports. This also was the case for professional groups, people, & events. References to Ana Kournikova, Tiger Woods, and Henry Cooper referred to people of current interest at the time or to key figures in the history of a sport. The equipment & facilities category queries reflected different hardware or specialized needs of each sport. Notably, sports such as wrestling, soccer, and boxing, which are not known for extensive equipment needs, did not contain high-frequency keywords related to

DOI: 10.1002/asi

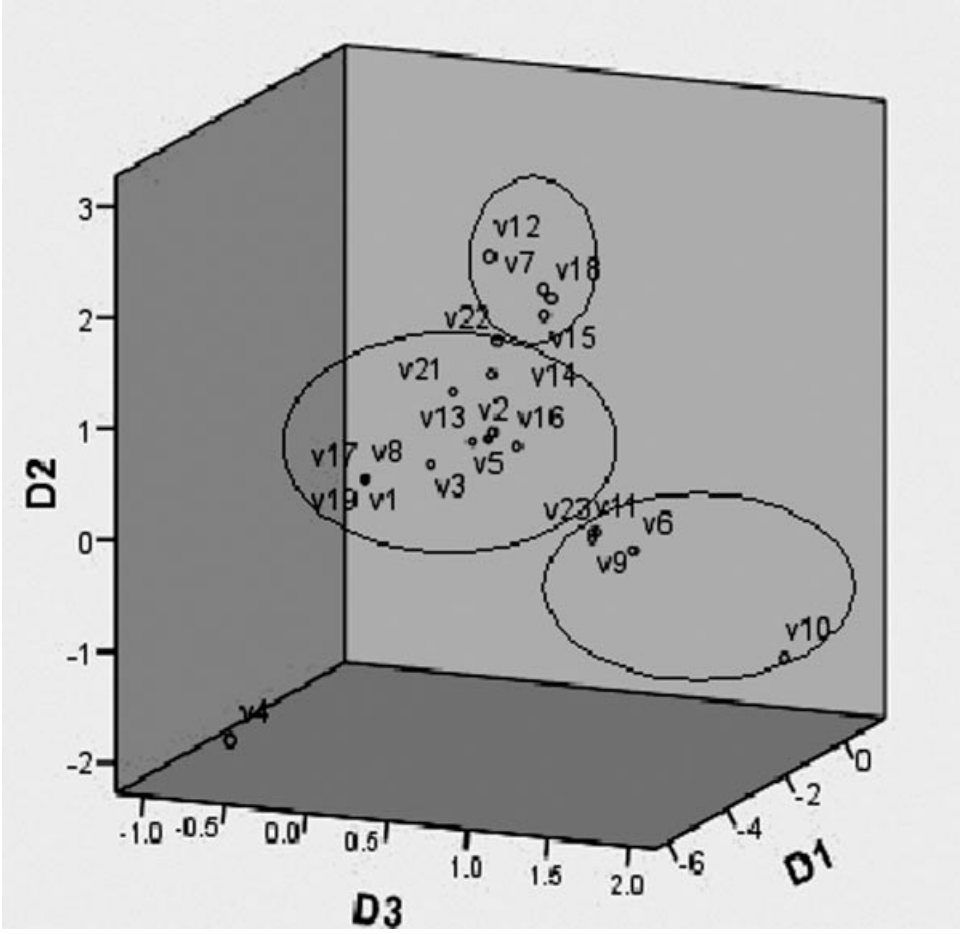

FIG. 9. Visual display for focus keyword *bowling*.

equipment. The computer & IT category reflected queries that combined sports information needs with technological themes such as the downloading of sports media or playing online sports games. Associated keywords such as *demo*, *computer*, *clip*, *onlinesports*, *online*, *Web*, *network*, *download*, *screen saver*, and *virus*, which are related to this theme, occurred in six focus keywords groups. Finally, six focus keyword groups contained query keywords related to gender, and more specifically, females in all but one case (men). The investigators observed that five of the six focus keyword groups also contained terminology of a sexual nature, although those keywords did not always group into the same cluster as the gender-based keywords. Conversely, sex-related keywords did not appear on associated keyword lists for focus keywords where no reference was made to females.

Finally, note that several different sports co-occurred in queries. For example, when the keyword *football* was used as the focus keyword, both *basketball* and *hockey* appeared in the related keyword group. When the focus keyword *sports* was employed, *football* and *swim* were associated with it. For the focus keyword *hockey*, both *basketball* and *football* were found in its associated keyword set. *Tennis* and *golf* also were observed to co-occur. This indicates that not all sports-related queries are specific to individual sports or deal with sports-related topics such as memorabilia, which can cover a number of different sports.

Search-Effectiveness Analysis

To test whether the relevant terms generated from the users' queries would help users improve search effectiveness, an experimental study was conducted. Eleven search tasks were designed based on the 11 investigated focus keywords (*golf*, *sports*, *ski*, *football*, *wrestling*, *hockey*, *soccer*, *baseball*, *tennis*, *boxing*, and *bowling*). Each search task represented one to three information needs; participants developed their search queries based on the information needs. Participants were instructed to select an information need and to conduct two searches in Google. In one search, the participant translated the selected information need into a query without any assistance; in the other search, the participant formulated a query with the assistance of a search-term list generated from relevant queries in the transaction log. The list includes the related search terms to a search task, co-occurrences, and the corresponding term clusters. Participants could use the search-term list to form their queries during the search.

The search effectiveness is defined as follows. The top-20 hits (i.e., Web sites) from the returned search-results list

DOI: 10.1002/asi

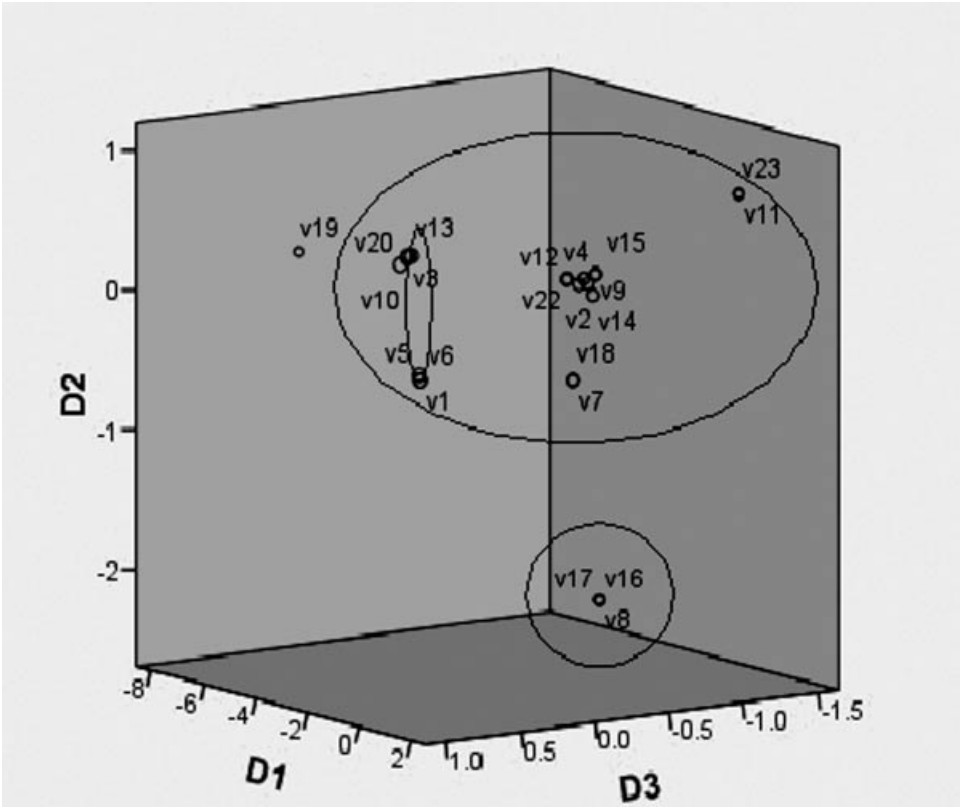

FIG. 10. Visual display for focus keyword *tennis*.

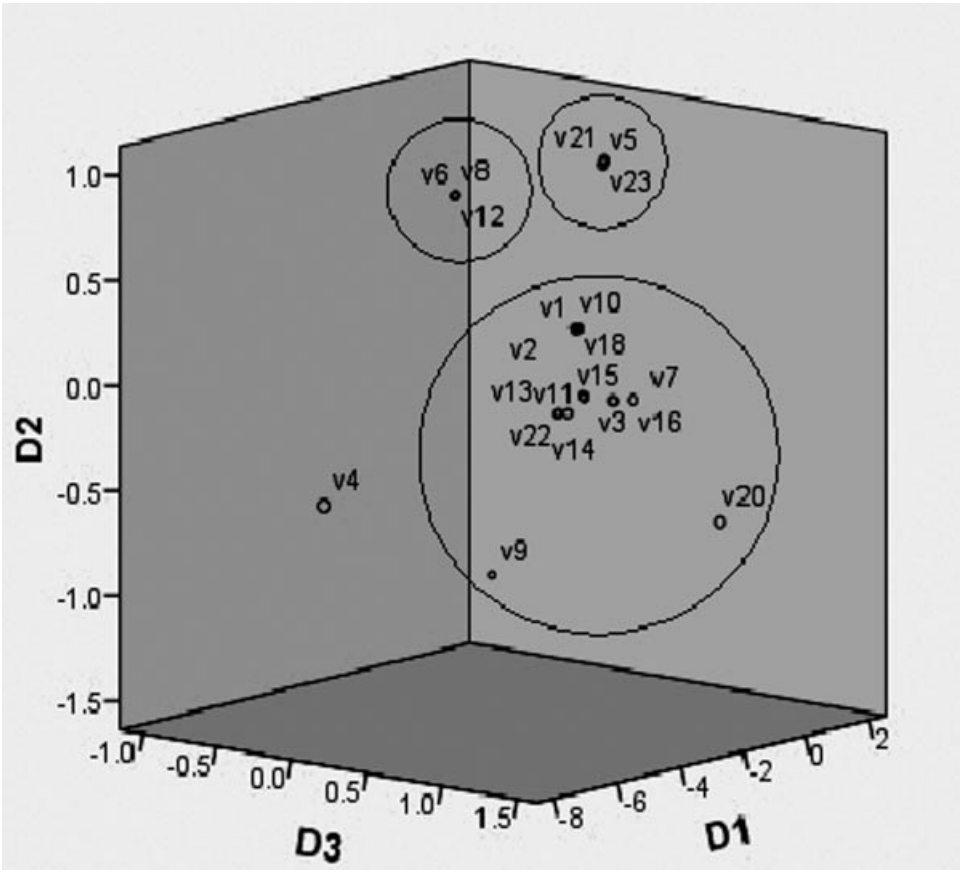

FIG. 11. Visual display for focus keyword *boxing*.

DOI: 10.1002/asi

TABLE 4. Summary of the focus keywords and topicality of associated terms.

	Education & training	Services	Geographic location	Professional groups, people, & events	Equipment & facilities	Computer & IT	Gender-related
Golf	x	x	x	x	x		x
Football	x	x	x	x	x	x	
Sports	x	x	x	x	x	x	x
Ski		x	x	x	x	x	
Wrestling	x	x	x	x		x	x
Hockey			x	x	x	x	
Baseball			x	x	x		
Soccer	x		x	x			x
Bowling	x		x	x	x	x	
Tennis	x		x	x	x		x
Boxing	x	x	x	x			x
Total	8	6	11	11	8	6	6

were examined (Sponsored Web sites were excluded.) Each of the 20 Web sites was assigned a relevance score (w_i) by the searcher, where $0 \leq w_i \leq 1$. The w_i values of 0, 0.5, and 1 corresponded to Web sites that were not relevant, partly relevant, and highly relevant, respectively. The effectiveness of a search was then calculated by averaging the relevance scores across the top-20 hits using the following equation.

$$Effectiveness = \frac{\sum_{i=1}^{20} w_i}{20} \quad (4)$$

After 11 search tasks were conducted and 44 searches were completed, the search-effectiveness results were tabulated. A t test was then used to determine whether the difference between the two groups of search performance was statistically significant. The mean and *SD* of the first group (search without assistance) were 0.610 and 0.160, respectively. The mean and *SD* of the second group (search with assistance) were 0.763 and 0.148, respectively. The t-test outcome was 3.30 ($df = 42$), with a p value of .002. This suggests that there is a significant difference in search effectiveness between the two groups. In other words, the searches that used search-term assistance by relying on previous query-term relationships outperformed the searches without the search-term assistance.

After the searches were completed, investigators interviewed the participants regarding their search experiences in the study. The participants commented that the search-term-assistance list did help them formulate their queries. The co-occurrence of a term, the neighboring terms of a term of interest in the alphabetic term list, and the clustered terms made a positive contribution in translating the information needs into more refined queries. It is no coincidence that search engines such as Google and Yahoo! have integrated past user queries in their search engine interface. When users enter a search query, the search engines automatically provide a list of relevant search terms to help the users. But the underlying methods supporting such term lists vary from system to system.

Discussion

This research demonstrates the feasibility of using hierarchical cluster analysis along with MDS for analyzing co-occurring query keywords related to specific themes extracted from query logs. By identifying those keywords that co-occur relatively more frequently with other keywords, coherent search-topic themes emerge from the analysis. Usually, each focus keyword produced three to five clusters. The largest cluster occupied the center of the visual space in the MDS mapping of the keywords.

One of the limitations of this study is that the investigated data were not extracted and generated from a recent query log. Unfortunately, due to data availability, the study was conducted with a relatively old dataset from the Excite search engine query log (1999). Findings reflect the search interests of the time. Over time, major players in sports change, sports events may never happen again, and new trends may appear. These can affect searchers' usage of keywords in queries. Despite these changes, the revealed fundamental information-seeking patterns such as the genre, basic search keywords (e.g., major sports leagues, events, equipment, and places), and sports keyword connections may stay the same. A longitudinal analysis with a more current dataset that applies the same methodology could demonstrate the longevity or ephemeral nature of some of the keyword associations. In addition, because the Excite data represent topics searched on a single day, the relationships among the keywords might vary seasonally. Still, the adopted methodology can be applied to any query log data. It lends itself best to environments where co-occurrence relationships are not sparse. In situations where this is the case, using cutoff values to eliminate sparsely co-occurring keywords is needed. As with other query log studies, the queries themselves cannot tell investigators why searchers searched for the topics or why they selected the keywords.

The analysis revealed several notable findings. A sports-related genre/category emerged from the analysis of the 11 selected focus keywords. The sports-related genre includes elements related to education and training;

DOI: 10.1002/asi

services; geographic locations; professional groups, people, and events; equipment and facilities; computer and IT; and gender-related topics. Among these, geographic locations; professional groups, people, and events; equipment and facilities; and education and training were the four most frequently used elements for all investigated focus keywords. Thus, they were important for sports. These genre elements were usually access points for sports-related searches. Keywords related to geographic location and professional groups, people, and events were commonly used in the users' queries. The phenomena indicate that users' searches were usually narrowed down to more specific queries by adding these keywords to the queries. It is not surprising that education and training was included in the genre because children and young people are frequent participants in a variety of sports. The identified computer and IT element in the genre confirms the claim that IT plays an important role in sports.

For each of the investigated focus keywords, a group of related keywords was identified, the clusters of the focus keyword were generated, and their relationships also were recognized. For instance, Tiger Woods, U.S. open tournament, and carts were highly associated with the focus keyword *golf*; and playoff, AFC, NFC, and FFL with the focus keyword *football*; and so on.

Note that in the visual MDS clustering-analysis results for all the investigated focus keywords, several relatively small, highly related, and well-defined clusters were identified in each of the resultant visual displays, such as the cluster {*open*, *U.S.*, *tournament*} for the focus keyword *golf*, and the cluster {*AFC*, *NFC*, *NFL*, *Sportsbook*} for focus keyword *football*. At the same time, there was a relatively large cluster that included a larger number of keywords in each of the visual displays with less obvious connections; for example, for the focus keyword *golf*, the cluster{*Arizona*, *beach*, *California*, *carts*, *Diego*, *equipment*, *greens*, *irons*, *Myrtle*, *packages*, *San*, *school*, *senior*, *state*, *travel*, *used*, *vacations*, *Washington*, *women*}. The relatively large clusters were confirmed in the two traditional clustering-analysis methods (Ward's method and the average linkage method) in most cases. The outcomes suggest that the relatively large and inclusive clusters were not a result of a specific clustering-analysis method. They truly reflect the connection among the involved keywords in the clusters. Unlike the keywords in the relatively small, highly related, and well-defined clusters, the keywords in the relatively large and inclusive clusters (e.g., *state*, *school*, *travel*, etc.) are more generic, and they may lose specific meanings if they are pulled out of their contexts. Therefore, they tended to appear together to form meaningful and specific queries. They have weaker, but more numerous, ties with other terms.

Both the information-visualization clustering method and the traditional clustering methods have their advantages and disadvantages. The information-visualization clustering method can give an intuitive, holistic, and straightforward global view of the clustered objects in the visual space. This global view is not only vivid but also easily understood by laypeople. It can effectively demonstrate multiple relationships among the objects and multiple relationships among identified clusters. Multiple connections between an object and other objects are crucial for a clustering analysis. The multiple connections are rarely reflected in a traditional clustering-analysis method. In addition, the visual clustering-analysis method can illustrate the extent to which an object (i.e., cluster) is associated with other objects (i.e., clusters). Defining the boundary for an emerging cluster is flexible in the visual space. It is not a problem if all objects are clearly spread out in the visual space; however, if the investigated objects are not scattered in a visual clustering environment, or the objects are mapped onto a relatively small area to create a high-density display, it is hard to unambiguously and explicitly define a boundary for a cluster.

On the other hand, the traditional clustering methods have unique strengths, although they do not possess the salient and prominent characteristics that a visual clustering-analysis method does. The most important feature of the traditional clustering methods is that they can clearly define emerging clusters for the objects; that is, they can effectively define the boundary for each identified cluster. This combined approach has been applied for many years in co-citation analysis studies (e.g., McCain, 1986). The aforementioned analysis suggests that combining visual-clustering analysis with hierarchical-clustering analysis can overcome the weaknesses of either method and achieve complementary clustering-analysis results. For instance, in Figure 1 (*golf*), Cluster C5 is located in the middle of the Clusters C1, C2, C3, and C4, and its cluster boundary is implicit. Without the use of the traditional clustering method, it would be difficult to define the boundary of the C5. Another example is in Figure 4 (*ski*), where the largest cluster, C3, was in the same situation.

In this study, the selected 11 focus keywords (*golf*, *football*, *sports*, *ski*, *hockey*, *wrestling*, *baseball*, *soccer*, *bowling*, *tennis*, and *boxing*) were representative and covered the most popular and major sports. The research method used also was applied to a more comprehensive and complicated customer-health-information area, where query keywords are more dynamic and intricate in nature. The findings in both the more general case (Zhang, Wolfram, Wang, Hong, & Gillis, 2008) and the more specific case (Zhang & Wolfram, 2009) were satisfactory.

Whether a keyword from the query dataset is selected as the focus keyword depends on multiple factors such as the nature and relevance of a selected query keyword to the study, its raw frequency in the query log, and a sound visual clustering analysis result based on the selected keyword and its associated keywords. The keyword *basketball* was certainly relevant to the study, and it maintained a reasonably high raw frequency (449) in the query log. For these reasons, it was initially selected as a focus keyword candidate; however, *basketball* and its associated keywords did not generate a sound visual-clustering result. In other words, the resultant stress values for basketball were greater than 0.10 (the acceptable stress value). We tried every possible means to minimize the resultant stress value for the focus keyword *basketball*, but failed to produce a satisfactory one. These measures included

DOI: 10.1002/asi

adjusting the Minkowski parameter, adapting the cutoff point for the co-occurrence of a focus keyword and an associated keyword (*CPCFKAK*), and modifying the cutoff point for total query keyword co-occurrence (*CPTQKC*). These measures affected the resultant stress values. This case suggests that a focus keyword with a high raw frequency in the query log does not guarantee a sound MDS visual display for the keyword.

The findings of this study not only provide a better understanding of sports-related information-seeking behavior but also have implications for sports-oriented system design and development. Thesauri, which usually include broad keywords, narrow keywords, synonyms, antonyms, and related keywords, are widely used to enhance information-retrieval effectiveness. Although a thesaurus is designed, developed, and maintained by information professionals, end-users would definitely benefit from a user-centered thesaurus that incorporates end-user vocabularies. It is evident that because the investigated keywords directly came from end-users, the identified keywords and the relationships among related keywords/clusters truly reflected end-users' habits and preferences in formulating queries. Therefore, the findings of this study can be used to enrich a sports-related thesaurus by adding new user-oriented entries to the thesaurus, establishing new user-oriented keyword connections, and including new user-oriented, relevant keywords.

Subject directories and indexes also are widely employed in information systems (especially in growing Web portals) as effective information-browsing tools guiding users to navigate the subject hierarchy. The revealed sports-related genre/category system in this study can be utilized to develop a user-centered facet subject directory for a sports-oriented information system or Web portal. The genre/category system was generalized and characterized for major sports. The items in the genre/category system were frequently the access points for sports-related searches. The genre also can be used to optimize the contents of a sports-oriented Web portal. Each element in the genre is associated with a frequency, which indirectly indicates the significance of the element. The contents of a sports-oriented portal can be classified and arranged based on the genre and element significance. The important contents can be emphasized and highlighted in the salient position of the portal for users' convenient access.

Query-search capability is fundamental and pivotal for an information system. A sports-oriented information system is no exception. By nature, search is an iterative process. Queries are modified towards a more relevant direction during the search process. As a result, query modification and query expansion are relatively complicated processes. If a group of related keywords is provided to users, it would assist users in reformulating their queries. In this study, the keywords were related to some extent if they appeared in the same cluster. The identified keywords and relevant relationships in a cluster can be used for query modification and query expansion. Keywords in a cluster are integrated as a group into the system. When a search query is submitted to a sports-oriented information system, a group of keywords related to the query keywords can be displayed automatically to the searcher so that he or she can select the suitable keywords from the list to modify and expand the original query. This approach can make a sports-oriented information-retrieval system more user friendly and effective.

The comparison of the results from the average linkage clustering method to those of Ward's clustering method showed that of the 11 focus keywords, the 4 keywords *football*, *sports*, *bowling*, and *boxing* yielded exactly the same results, and the 3 keywords *golf*, *baseball*, and *soccer* yielded almost the same results with the exception of one or two keywords. However, the numbers of clusters created by the two traditional clustering methods were not equivalent for the two focus keywords, *hockey* and *tennis*. Aside from a new cluster, which emerged in two cases, the resulting clusters were the same. There were significant differences between the two traditional methods for the two keywords *wrestling* and *ski*. Their visual displays clearly show that dense keyword distributions make it difficult to distinguish the clusters. This implies that if there are significant differences between the two traditional methods, the visual-display results can truly help people understand the nature of the difference.

Conclusion

This study examined a public search engine query log file to gain insights into the information-seeking behavior of Internet searchers on the most popular sports-related topics entered by users. Based on focus keyword raw frequencies in a master query keyword file and their popularity, 11 sports keywords were selected for the study. A method was developed for identifying appropriate search keywords that co-occurred with the focus keywords and would allow for meaningful relationships to be identified among the keywords. Multidimensional scaling and traditional hierarchical clustering were applied to focus keywords and their relevant keywords to identify groups of related keywords. The results of a hierarchical cluster analysis were integrated in the corresponding MDS visual-display contexts to provide a clearer picture of how query keywords related to one another. The calculated MDS stress and RSQ values of these investigated focus keywords demonstrated that each of the visual analyses yielded satisfactory and sound results.

If a sports-related keyword is used in a query, its related keywords in the cluster can be used to either expand or revise the query. An experimental study was conducted to determine whether there was a significant difference between the searches that used search-term assistance by relying on previous query-term relationships and the searches without the search-term assistance in terms of search effectiveness. The experimental results revealed that the searches that used search-term assistance outperformed the searches without the search-term assistance.

These findings can help researchers better understand online sports-information-seeking behavior. For each focus keyword, three to five related keyword clusters with a large central cluster were identified. Keywords in an identified

DOI: 10.1002/asi

cluster were semantically associated. Although the focus keywords were different and the corresponding relevant keyword sets varied substantially in size and content, the relevant keywords shared some commonalities. These relevant keywords were categorized into seven distinct categories (education & training; services; geographic location; professional groups, people, & events; equipment & facilities; computer & IT; and gender-related topics). In other words, the most relevant keywords to a focus keyword can be grouped into one of these distinct categories. The findings also revealed relevant subjects and subtopics of a given sport. Further research will continue to refine the developed method and apply it to other query log datasets and topics.

Acknowledgment

We thank *Excite@home*, the University of Tennessee, and *HealthLink* for access to the query log data. This research was funded by Institute of Museum and Library Services National Leadership Research Grant LG-06-05-0100-05). Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the Institute of Museum and Library Services.

References

Baeza-Yates, R., Dupret, G., & Velasco, J. (2007, May). A study of mobile search queries in Japan. Paper presented at the 16th International World Wide Web Conference, Workshop on Query Log Analysis: Social and Technological Challenges, Banff, Alberta, Canada. Retrieved April 29, 2009, from http://www2007.org/workshops/paper_50.pdf

Beeferman, D., & Berger, A. (2000). Agglomerative clustering of a search engine query log. In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 407–416), Boston. New York: ACM Press.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Frieder, O., & Grossman, D. (2007). Temporal analysis of a very large topically categorized Web query log. Journal of the American Society for Information Science and Technology, 58(2), 166–178.

Borish, L.J. (1992). The sporting past in American history. History of Sport, Recreation, and Leisure, 7(1). Retrieved April 29, 2009, from http://www.oah.org/pubs/magazine/sport/borish.html

Chalmers, M., & Chitson, P. (1992). Bead: Explorations in information visualization. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 330–337). New York: ACM Press.

Cooper, M.D. (2001). Usage patterns of a Web-based library catalog. Journal of the American Society for Information Science and Technology, 52, 137–148.

Croft, W.B., Cook, R., & Wilder, D. (1995). Providing government information on the Internet: Experiences with THOMAS. In Proceedings of the Second International Conference on Theory and Practice of Digital Libraries (pp. 19–24). Retrieved April 29, 2009, from http://csdl.tamu.edu/DL95/papers/croft/croft.html

Dutton, W., & Helpser, E. (2007). Oxford internet survey: The internet in Britain 2007. Oxford, UK: Oxford Internet Institute.

Ernest, D.J., Level, A.V., & Culbertson, M. (2005). Information seeking behavior for recreational activities and its implications for libraries. Reference Services Review, 33(1), 88–103.

Fowler, R.H., Fowler, W.A.L., & Wilson, B.A. (1991). Integrating query, thesaurus, and documents through a common visual representation. In Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information (pp. 142–151). New York: ACM Press.

Free, M., & Hughson, J. (2006). Common culture, commodity fetishism and the cultural contradictions of sport. International Journal of Cultural Studies, 9(1), 83–104.

Goodrum, A., & Spink, A. (2001). Image searching on the Excite Web search engine. Information Processing & Management, 37(2), 295–312.

Hakanen, E.A., & Wolfram, D. (1995). Citation relationships among international mass communication journals. Journal of Information Science, 22(3), 9–15.

Hallam-Baker, P.M., & Behlendorf, B. (2008). Extended log file format. Retrieved April 29, 2009, from http://www.w3.org/TR/WD-logfile

Hirsh, S.G. (1999). Children's relevance criteria and information seeking on electronic resources. Journal of the American Society for Information Science, 50(14), 1265–1283.

Hoelscher, C. (1998). How Internet experts search for information on the Web. In H. Maurer & R.G. Olson (Eds.), Proceedings of WebNet98—World Conference of the WWW, Internet & Intranet. Chesapeake, VA: AACE.

Huang, C., Chien, L., & Oyang, Y. (2003). Relevant term suggestion in interactive Web search based on contextual information in query session logs. Journal of the American Society for Information Science and Technology, 54(7), 638–649.

Jansen, B.J., Goodrum, A., & Spink, A. (2000). Searching for multimedia: An analysis of audio, video and image Web queries. World Wide Web: An International Journal, 3(4), 249–254.

Jansen, B.J., Jansen, K., & Spink, A. (2005). Using the Web to look for work: Implications for online job seeking and recruiting. Internet Research: Electronic Networking Applications and Policy, 15(1), 49–66.

Jansen, B.J., Spink, A., & Koshman, S. (2007). Web searcher interaction with the Dogpile.com metasearch engine. Journal of the American Society for Information Science and Technology, 58(5), 744–755.

Jones, S., Cunningham, S.J., & McNab, R.J. (1998). An analysis of usage of a digital library. In Proceedings of the Second European Conference on Digital Libraries (pp. 261–277). New York: ACM Press.

Joshi, K.P., Joshi, A., Yesha, Y., & Krishnapuram, R. (1999). Warehousing and mining Web logs. In Proceedings of the Second International Workshop on Web Information and Data Management (pp. 63–68). New York: ACM Press.

Kelly, D. (2006). Measuring online information seeking context, Part 2: Findings and discussion. Journal of the American Society for Information Science, 57(14), 1862–1874.

Kohonen, T. (2001). Self-organizing maps. Springer Series in Information Sciences, 30. New York: Springer.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela. A. (2000). Self organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3), 574–585.

Korfhage, R.R. (1997). Information storage and retrieval. New York: Wiley.

Koshman, S., Spink, A., & Jansen, B.J. (2006). Web searching on the Vivisimo search engine. Journal of the American Society for Information Science and Technology, 57(14), 1875–1887.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, 1–27.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: A numerical method. Psychometrika, 29, 115–129.

Mahoui, M., & Cunningham, S.J. (2000). A comparative transaction log analysis of two computing collections. In Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (pp. 418–423). London: Springer-Verlag.

McCain, K.W. (1986). Cocited author mapping as a valid representation of intellectual structure. Journal of the American Society for Information Science, 37(3), 111–122.

Park, S., Lee, J.H., & Bae, H.J. (2005). End user searching: A Web log analysis of NAVER, a Korean Web search engine. Library & Information Science Research, 27(2), 203–221.

Pew Internet and American Life Project. (2005). Teens and technology: Youth are leading the transition to a fully wired and mobile nation. Washington, DC: Author.

Rieh, S.Y., & Xie, H.I. (2006). Analysis of multiple query reformulations on the Web: The interactive information retrieval context. Information Processing & Management, 42(3), 751–768.

Robertson, G., Card, S.K., & Mackinlay, J.D. (1989). The cognitive coprocessor architecture for interactive user interfaces. In Proceedings of the Second Annual ACM SIGGRAPH Symposium on User Interface Software and Technology (pp. 10–18). New York: ACM Press.

Ross, N.C.M., & Wolfram, D. (2000). End-user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. Journal of the American Society for Information Science, 51(10), 949–958.

Schvaneveldt, R.W., Durso, F.T., & Dearholt, D.W. (1989). Networking in proximity data. In G. Bower (Ed.). The psychology of learning and motivation (Vol. 24, pp. 249–284). San Diego, CA: Academic Press.

Shi, X., & Yang, C.C. (2007). Mining related queries from Web search engine query logs using an improved association rule mining model. Journal of the American Society for Information Science and Technology, 58(12), 1871–1883.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large Web search engine query log. SIGIR Forum, 33(1), 6–12.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between documents. Journal of the American Society for Information Science, 24, 265–269.

Small, H., & Garfield, E. (1985). The geography of science: Disciplinary and national mapping. Journal of Information Science, 11, 147–159.

Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. Computer Magazine, 35(3), 107–109.

Spink A., Ozmutlu, H.C., & Lorence, D.P. (2004). Web searching for sexual information: An exploratory study. Information Processing & Management, 40(1), 113–123.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology, 52(3), 226–234.

Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D.P., Ozmutlu, S., & Ozmutlu, H.C. (2004). A study of medical and health queries to Web search engines. Health Information and Libraries Journal, 21(1), 44–51.

SportsBusiness Daily. (2001). Study shows online behavior of sports fans "predictable." Retrieved April 29, 2009, from http://www.sportsbusinessdaily.com/article/61189

Stappers, P.J., & Pasman, G. (1999). Exploring a database through interactive visualised similarity scaling. In Proceedings of CHI '99 Extended Abstracts on Human Factors in Computing Systems (pp. 184–185). New York: ACM Press.

Thelwall, M. (2002). An initial exploration of the link relationship between UK university Web sites. ASLIB Proceedings, 54(2), 118–126.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. Psychometrika, 17, 401–419.

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.

Wang, P., Berry, M.W., & Yang, Y. (2003). Mining longitudinal Web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54(8), 743–758.

Wen, J., Nie, J., & Zhang, H. (2001). Clustering user queries of a search engine. In Proceedings of the 10th International Conference on World Wide Web (pp. 162–168). New York: ACM Press.

White, H., & McCain, K. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972–1995. Journal of the American Society for Information Science, 49(4), 327–355.

Whittle, M., Eaglestone, B., Ford, N., Gillet, V.J., & Madden, A. (2007). Data mining of search engine logs. Journal of the American Society for Information Science and Technology, 58(14), 2382–2400.

Yi, K., Beheshti, J., Cole, C., Leide, J.E., & Large, A. (2006). User search behavior of domain-specific information retrieval systems: An analysis of the query logs from PsycINFO and ABC-Clio's Historical Abstracts/America: History and Life: Research Articles. Journal of the American Society for Information Science and Technology, 57(9), 1208–1220.

York, J., Bohn, S., Pennock, K., & Lantrip, D. (1995). Clustering and dimensionality reduction in SPIRE. In AIPA Steering Group (Eds.), Proceedings of the Symposium on Advanced Intelligence Processing and Analysis (p. 73). Washington, DC: Office of Research and Development.

Zeki, S. (1992). The visual image in mind and brain. Scientific American, 267(3), 69–76.

Zhang, J. (2001). TOFIR: A tool of facilitating information retrieval—Introduce a visual retrieval model. Information Processing & Management, 37(4), 639–657.

Zhang, J. (2008). Visualization for information retrieval. Berlin, Germany: Springer.

Zhang, J., & Korfhage, R.R. (1999). DARE: Distance and angle retrieval environment: A tale of the two measures. Journal of the American Society for Information Science, 50(9), 779–787.

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

Zhang, J., Wolfram, D., Wang, P., Hong, H., & 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.

DOI: 10.1002/asi