How Many Ways to Use CiteSpace? A Study of User Interactive Events Over 14 Months

Qing Ping,Jiangen He, and Chaomei Chen
College of Computing & Informatics, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104.
E-mail:qp27@drexel.edu; jh3328@drexel.edu; cc345@drexel.edu

Using visual analytic systems effectively may incur a steep learning curve for users, especially for those who have little prior knowledge of either using the tool or accomplishing analytic tasks. How do users deal with a steep learning curve over time? Are there particularly problematic aspects of an analytic process? In this article we investigate these questions through an integrative study of the use of CiteSpace—a visual analytic tool for finding trends and patterns in scientific literature. In particular, we analyze millions of interactive events in logs generated by users worldwide over a 14-month period. The key findings are: (i) three levels of proficiency are identified, namely, level 1: low proficiency, level 2: intermediate proficiency, and level 3: high proficiency, and (ii) behavioral patterns at level 3 are resulted from a more engaging interaction with the system, involving a wider variety of events and being characterized by longer state transition paths, whereas behavioral patterns at levels 1 and 2 seem to focus on learning how to use the tool. This study contributes to the development and evaluation of visual analytic systems in realistic settings and provides a valuable addition to the study of interactive visual analytic processes.

Introduction

Recent years have witnessed an increasing popularity of visual analytics, including commercial products, such as Tableau[1] (Stolte, Tang, & Hanrahan, 2002) and GeoTime[2] (Eccles, Kapler, Harper, & Wright, 2008), freely available software, such as Many Eyes (Viegas, Wattenberg, Van Ham, Kriss, & McKeon, 2007) and Jigsaw[3] (Stasko, Görg, & Liu, 2008), and general-purpose visualization tools, such as D3.js[4] (Bostock & Heer, 2009). Some of the visual analytics applications, such as CiteSpace[5] (Chen, 2006), VOSViewer[6] (Van Eck & Waltman, 2010), and Action Science Explorer[7] (Cobo, López-Herrera, Herrera-Viedma, & Herrera, 2011), are specifically designed for understanding a knowledge domain from a large volume of scientific literature.

Longitudinal studies of how users interact with visual analytic applications in real-world settings are relatively rare. Earlier studies primarily focused on usage patterns in a laboratory setting (Allendoerfer et al., 2005; Gove, Dunne, Shneiderman, Klavans, & Dorr, 2011; Heer, Card, & Landay, 2005; Yuan, Chen, Zhang, Avery, & Xu, 2013) or on observations and user logs in naturalistic settings (Bostock & Heer, 2009; Harris & Butterworth, 2012; Hindus, Ackerman, Mainwaring, & Starr, 1996; Liu, Liu, & Wang, 2010; Schmid, 2012; Viegas et al., 2007). Laboratory studies typically examine user activities of a small group of participants over a short period of time. User behavioral patterns, especially when adapting to a new technology, however, may take much longer to emerge and become detectable. Moreover, observations made in laboratory settings can be intrusive and interfere with the normal workflow. Naturalistic observations, such as video recording and self-diary, may impose an extra cognitive burden on users. Log analysis, typically data-driven and nonintrusive, may reveal emergent patterns of user activities. On the other hand, the approach is limited in terms of its ability to identify underlying perceptual and cognitive processes behind behavioral patterns. Therefore, using an integrative approach, that is, a quantitative study of logs of interactive events triangulated with qualitative analysis, is more effective than using a single-method approach.

CiteSpace is a freely available visual analytic application for users to analyze and visualize a knowledge domain based on relevant publications (Chen, 2004, 2006). CiteSpace has undergone several usability studies, for example, heuristic evaluation and usability tests with a perceptual-cognitive taxonomy mapping (Synnestvedt & Chen, 2005), cognitive walk-through evaluation (Allendoerfer et al., 2005), and a study of the effects of users' domain knowledge on their performance and perception of the system (Yuan et al., 2013).

[1]http://www.tableau.com/
[2]http://geotime.com/
[3]http://www.cc.gatech.edu/gvu/ii/jigsaw/
[4]http://d3js.org
[5]http://cluster.cis.drexel.edu/~cchen/citespace/
[6]http://www.vosviewer.com/Home
[7]http://www.cs.umd.edu/hcil/ase/#software

Received June 26, 2015; revised April 21, 2016; accepted May 23, 2016

 DOI: 10.1002/asi.23770

The present study expands both the breadth and depth of previous studies of how users interact with a visual analytic system in real-world settings. We study the use log of a much larger population than those in previous studies over a much longer period of time. Our study is nonintrusive in the context of users' own settings and questions posted by a diverse population of users on a public forum. The long-term goal of the research is to understand how users navigate through a space of various potential paths such that one can optimize the design of visual analytic systems such as CiteSpace. A better understanding of users' procedural patterns will help us to improve the design of application visual analytic system, and enable novice users to learn from users who are more proficient with the tool, more knowledgeable of a subject domain, or both. More specifically, the present study attempts to achieve the goal through the following steps: First, we characterize user behaviors in terms of groups similar paths of interaction; second, we profile interactive patterns of each group with multiple facets or metrics; third, we focus on patterns of adaptation, namely, how likely each group of users would switch to a new version. Finally, we underline error-prone areas by matching behavioral patterns with issues raised by users on a public forum, and propose design implications based on lessons learned. All these analyses are driven by the large-scale user log data and thousands of user feedback and questions posted on a blog in Chinese and answers provided by the third author. The blog data are analyzed and cross-referenced with the user event log data to identify areas for improvement.

In summary, we address four research questions in the present study:

1. Is it feasible to cluster use sessions based on behavioral patterns in CiteSpace?
2. What interactive patterns characterize each user group?
3. What adaptation patterns are associated with each user group?
4. Which parts of the analytic process of using CiteSpace are error-prone and what design improvements can be made?

Related Work

In this section we review two areas of related work: First, we present our study in the context of existing studies of the use of visual analytic systems. Next, we summarize various methods and techniques for clustering and profiling users and sessions in log analyses.

Usability Studies in Visual Analytics

Research of users' behavioral patterns in visual analytic applications is relatively rare, in part because the majority of visual analytic applications are still relatively young. Most previous studies included a user study as part of the introduction of a system. We divide the existing studies into laboratory-based usability studies and others, including field studies, although other divisions may be possible.

Laboratory usability studies. Usability studies in a laboratory setting typically investigate participants' performance in completing a series of tasks with the system. For example, in a study of Prefuse, an open-source visualization toolkit, eight participants were recruited to perform three programming tasks, which enabled the researchers to identify design and naming issues in Prefuse (Heer et al., 2005). For Action Science Explorer, a knowledge domain visualization application (KDViz), four graduate students participated in a usability study to perform both predefined and user-defined tasks in a "thinking loud" manner, which revealed both users' tendencies in using the system and design issues in the system (Gove et al., 2011). For Jigsaw, a text visual analytic application, 16 graduate students participated in a laboratory experiment to perform a simulated intelligence analysis task using Jigsaw, which enabled the developers of Jigsaw to identify the characteristics of a sense-making process and their design implications (Kang, Gorg, & Stasko, 2011). For the CiteSpace system, also a KDViz application, three user studies were conducted. One study utilized a modified "cognitive walk-through" method with six graduate students to evaluate the design of CiteSpace, which led to the identification of five design issues (Allendoerfer et al., 2005). In another study, three evaluators conducted a heuristic evaluation and a usability test to evaluate the design of CiteSpace in terms of a cognitive-perceptual taxonomy (Synnestvedt & Chen, 2005). The third study was conducted as a laboratory experiment with 32 participants, and it found that the users' performance and perception were significantly affected by their levels of domain knowledge (Yuan et al., 2013).

Laboratory-based usability studies examine how participants interact with a system. However, laboratory studies also have constraints: users' performance can be only observed during a short period of time; many studies tend to cover a small subset of a usually much larger set of functions participated by a small portion of the entire user population. In contrast, in a real-world setting, user behavioral patterns may take much longer to emerge. Similarly, given a complex visual analytic application with many functions, it is often difficult to examine all the functions within a single usability test. Furthermore, users' levels of knowledge vary greatly, especially between novices and experts. Finally, laboratory-based experiments can be intrusive (Godoy & Amandi, 2005), which may distort participants' performance.

Naturalistic observations. Researchers have studied user behavior of using visual analytic applications in naturalistic settings. Naturalistic observation methods include self-report diary records, on-site observation, videotaping (Harris & Butterworth, 2012), as well as log recording. Among these methods, log analysis and self-report are the most often used for users of visual analytic applications. For example, the designers of Protovis, an open-source visualization library,

DOI: 10.1002/asi

revised their system based on user feedback they gathered (Bostock & Heer, 2009). The designers of IBM's Many Eyes, a web-based visual analytic application, collected not only user feedback but also the percentages of the usage of different visualization components from over 1,000 web users (Viegas et al., 2007). Tableau, a commercial visual analytic product, collected both informal feedback from its own employees and logs of user activities on a subset of the system, which led to a timeline view of the system usage and a comparison of the usage of different functions (Mackinlay, Hanrahan, & Stolte, 2007). The VisTrail application, a comprehensive visualization provenance management infrastructure, provides a mechanism of history management to automatically trace user's visualization operations for analysis (Bavoil et al., 2005; Callahan et al., 2006). The annual Visual Analytics Science and Technology (VAST) challenge is an annual contest among visual analytics researchers, developers, and designers to apply their best tools and techniques to invented problems (Cook, Grinstein, & Whiting, 2014). Another example is the study of Jigsaw, in which the researchers reflected upon lessons learned from their own participation in the VAST contests (Görg, Liu, & Stasko, 2013).

Besides log analysis and self-report, other naturalistic methods are often used in combination with data-driven methods to analyze user behaviors. For example, diary recording was used as a supplement of log analysis to provide contextual information to understand user behavior in a job-searching application (Liu et al., 2010). A questionnaire was used to collect user diary as well as usage logs for usage analysis (Schmid, 2012). Videotaping, direct observation, and interviews are coupled with log recording in a 2-month field study to analyze the usability of an audio-only media space (Hindus et al., 1996).

Naturalistic studies are nonintrusive from a user's point of view. However, they also have limitations. Users in diary recording often suffer from distraction as well as extra cognitive burden to keep a diary during their normal workflow (Rieman, 1993). During videotaping, users could block the cameras unintentionally and are influenced by being videotaped (Arhippainen & Tähti, 2003). Moreover, findings of field studies may be hard to generalize to a broader context (Stephens, 1982). Log analysis studies have the potential to analyze a larger number of participants over a longer period of time.

The present study aims to bridge the gap by investigating users' visual analytic behavior based on a large user population over a relatively long period of time based on user logs collected nonintrusively. Our study cross-examines findings resulting from log analyses with a content analysis of user feedback from a blog on CiteSpace. Major characteristics of the present study are summarized in Table 1 in the context of other relevant studies. For instance, the present study has the longest window of observation (14.6 months) and a large international population of users across diverse levels of proficiency (18,049 IP addresses). Our study analyses both interactive and adaptive behavioral patterns of users.

Log Analysis for User/Session Clustering and Profiling

Two broad categories of log analyses are commonly seen, namely, web content analyses and user/session pattern analyses. The latter is more relevant in the present context.

Web content log analysis. web content analysis studies web server logs in an attempt to develop a good understanding of how a website has been used and how it can be improved using techniques such as query expansion, document clustering, and information extraction (Fuxman, Tsaparas, Achan, & Agrawal, 2008; Lu, Wilbur, McEntyre, Iskhakov, & Szilagyi, 2009; Zhang & Nasraoui, 2006; Zhao et al., 2006).

User/session pattern analysis. user/session pattern analysis aims to develop user models or profiles based on usage logs (Fischer, 2001). More specifically, in the context of log analysis, a variety of user characteristics have been profiled, ranging from descriptive statistics of user activities, to classifications of users' intentions and needs, to clustering of user behavioral patterns. Descriptive statistics of user activities include the length of a query and its variability (Kamvar, Kellar, Patel, & Xu, 2009), content and user revisit rates (Kumar & Tomkins, 2010), query categorization, patterns of search, clickstreams on different types of devices (Y. Song, Ma, Wang, & Wang, 2013), query categorization, and reformulation (Dogan, Murray, Névéol, & Lu, 2009).

For classification of users' intention or needs based on usage logs (Lee, Liu, & Cho, 2005), classified user goals (navigational versus informational) in web search based on features of user-click behavior and anchor-link distribution. Hu, Zhang, Chen, Wang, and Yang (2011) adds a third component, the user intention hypothesis, into the traditional click model that jointly estimates click-rate and position-bias, which achieves significantly better performance than click-models under only examination hypothesis. Cao, Jiang, Pei, Chen, and Li (2009), models user's search intent as the state in the variable-length hidden Markov model (vlHMM) based on past queries and clickstreams.

For user behavioral pattern clustering, users can be represented by either a nondependent model or matrix or a dependency model, such as a path, a tree, or a graph. The resulting clusters are a group of users who have visited webpages in a similar way. For nondependent representation, Yan, Jacobsen, Garcia-Molina, and Dayal (1996) represented users with a vector of frequencies over the webpages visited. Xie and Phoha (2001) represented users as a set of sessions paired with the basic probability assignment (originated from the Dempster-Shafer's theory) over all the sessions of the user. Song and Shepperd (2006) represented users and webpages with a user by webpage matrix with the element being the number of hits and clustered users and webpages simultaneously. Wan, Jönsson, Wang, Li, and Yang (2012) represented users with a vector space model of visited URL segments (random indexing) to obtain the latent contexts of user interests in webpages. For representation of dependency model, Shahabi, Zarkesh, Adibi, and Shah (1997) represented

DOI: 10.1002/asi

TABLE 1. Comparison of user studies of visual analytic applications.

System	Reference	#Users	Type of Users	Duration	Type of analysis	Knowledge required for users D-Domain S-System P-Programming
Many Eyes	(Viegas et al., 2007)	1,463	Users on the web	2 months	Basic statistics	D
Tableau	(Mackinlay et al., 2007)	17	Employees; users	11 months	Basic statistics; timeline view of logs	D/S
Prefuse	(Heer et al., 2005)	8	Programmers; CS students; IV experts	~2 hours	Usability study, qualitative	D/S/P
Action Science Explorer	(Gove et al., 2011)	4	2 PhDs; 2 graduate students	2.5 hours	Usability study, qualitative	D/S
VisTrail	(Silva, Anderson, Santos, & Freire, 2011)	30	30 students taking the Scientific Visualization course	2 semesters	Manual classification of tasks; visualization of version tree of users.	D/S
Jigsaw	(Kang et al., 2011)	16	graduate students	~3 hours	In-laboratory experiment; interview	D/S
	(Görg et al., 2013)	1	Developers themselves; other users	3 months for 3 years	Observation; self-reflection	D/S
CiteSpace	(Yuan et al., 2013)	32	16 graduate students and 16 undergraduates	~2 hours	In-laboratory experiment	D/S
	(Allendoerfer et al., 2005)	6	Graduates	2 hours	Cognitive walkthrough	D/S
	(Synnestvedt & Chen, 2005)	3	Medical researchers	2 hours	Heuristic evaluation; usability study	D/S
	Present study	18,049	the entire CiteSpace user population	14.6 months	User modeling and profiling; user adaptation pattern analysis; error analysis	D/S

sessions as paths of different lengths and calculated path similarity by calculating unions of all subpaths with the space of path features. Later, they represented sessions with a feature-matrices model, independent of cardinality (Shahabi & Banaei-Kashani, 2003). Huang, Ng, Cheung, Ng, and Ching (2001) represented use sessions as a cube model of session, access sequence order, and attribute of each access (webpage). Banerjee and Ghosh (2001) represented a session as a sequence of webpages and used the longest common sequence as the similarity measure to cluster users. Lin and Wilbur (2009) represents user actions as a sequence of symbols from a finite set of alphabets and used language modeling to cluster the sequences. Chen, Bhowmick, and Nejdl (2009) extracted frequent subtrees in the web session trees as FRACTURE to represent the evolutionary patterns of user behavior. Bayir, Toroslu, Cosar, and Fidan (2009) proposed a Smart-SRA algorithm to construct the graph of sessions, which is a set of paths traversed in the web graph. Dimopoulos, Makris, Panagis, Theodoridis, and Tsakalidis (2010) represented sessions with weighted suffix trees generated from weighted web access sequences. The representation of user as a nondependent model will computationally save time for later calculation of similarities between user using different distance metrics, at the cost of losing the information of sequence order. On the other hand, the representation of dependency models is able to capture the sequential information of user access at the cost of more complex representation and calculation of similarities between users. In the present study, we use nondependent representation, that is, vectors of frequencies over all events for session clustering.

User groups can be identified by various clustering algorithms, such as the leader algorithm (Yan et al., 1996), K-means (Shahabi et al., 1997; Wan et al., 2012; Xu & Liu, 2010), K-modes (Huang et al., 2001), dynamic clustering (Shahabi & Banaei-Kashani, 2003), association rule mining (Bayir et al., 2009; Mobasher, Cooley, & Srivastava, 1999), and fuzzy clustering (Gandy, Rahimi, & Gupta, 2005; Nasraoui, Frigui, Joshi, & Krishnapuram, 1999; Suryavanshi, Shiri, & Mudur, 2005). Other research also approached the problem by performing min-cut graph partitioning on a similarity graph of paths (Banerjee & Ghosh, 2001). Ling Chen et al. (2009) compared different clustering methods, that is, partitioning methods, agglomerative methods, and graph methods, and found that partitioning methods were preferable on their FRACTURE similarity among the three methodologies (L. Chen et al., 2009). Lastly, taking user actions as

DOI: 10.1002/asi

a sequence of symbols, n-gram language model was also used to find underlying contexts (patterns) of user actions (Lin & Wilbur, 2009). The choice of clustering methods is interrelated with the representation of the users. In the present study, we chose to use hierarchical clustering on the vectors of sessions that yields satisfactory results; other clustering methods can also be explored to further improve the clustering results in the future.

These studies have explored many techniques for analyzing web logs and clustering sessions. There are three basic steps: identify session, measure similarities (or dissimilarities) between session, and cluster session and generate session profiles. Our study differs from those described earlier in that instead of representing sessions as a sequence of webpages visited by a user, we represent sessions as a vector of events for clustering, and then later profile each cluster with a chain of events, that is, a state transition path in a state space defined by the types of events. We utilize a relatively simple representation of a session, for example, a frequency distribution of different events. To calculate the dissimilarity between sessions, the Kullback–Leibler divergence is used. Finally, we conduct agglomerative hierarchical clustering on similarities between sessions, and we generate profiles from aggregated state transition matrices within a cluster.

Methods

In this section we first describe the preprocessing of the event log data, including data collection, session identification, and IP-aggregated session. Second, we describe how we identify meaningful behavioral patterns by introducing session representations, a session similarity measure, and a session clustering method. Third, we answer the question concerning interactive patterns of session clusters, by comparing the usage and the coverage of various events, the reach of crucial events, and state transition patterns of each cluster. Fourth, we identify evolving patterns of clusters by modeling users' version adaptation behavior using survival analysis. Finally, we use content analysis and topic modeling to identify issues emerging from user feedback.

Data Preprocessing

In this section we describe the three steps of preprocessing: data collection, session identification, and IP-aggregated session.

Data collection. CiteSpace is a Graphic User Interface (GUI)-based visual analytic tool. A user would typically generate multiple events in several windows to accomplish a visual analytic task. These events are defined as atomic actions, such as a click on a menu, a selection from a drop-down box, or a specification of a number. These events are also semantically defined, for example, "labeling the cluster using LLR measure," which are unlike semantically free events (Gotz & Zhou, 2009). CiteSpace logs the occurrences of 81 types of events, each of which is given a distinct name. At the beginning of a user session, the system asks for the user's consent for it to collect event logs. As the user clicks through various controls on the user interface, the logging module records the corresponding events with timestamps, as shown in Figure 1. The logging module synchronizes the logs to a server. The schema of the logging data table stored in the server is shown in Table 2.

Session identification. A session is a series of activities related to each other at the conceptual level through close proximity in time (He, Göker, & Harper, 2002). We must identify sessions from a stream of logged events to understand user behavioral patterns. Previous studies consider a 30-minute idle time as a good cutoff point to identify sessions of web browsing (Srivastava, Cooley, Deshpande, & Tan, 2000). We follow similar heuristics:

- The event "version" marks the beginning of a session.
- The event "exit" marks the end of a session.
- If an event is followed by a 60-minute idle period, then the event marks the end of the current session.

IP-aggregated session. An IP-aggregated session is a collection of sessions generated by the same IP address during the data collection period. IP-aggregated session is the unit of clustering and pattern profiling in the present study. We decided to cluster IP-aggregated sessions instead of individual sessions for several reasons. First, clustering sessions (~250,000) is computationally challenging. Second, we observed that sessions were often dominated by one particular behavioral pattern, since this pattern generates far more sessions than the other patterns. As a result, clustering sessions directly would only reveal predominant behavioral patterns and omit other distinctive patterns. Therefore, we chose to perform clustering IP-aggregated sessions. When sessions are aggregated as a unit, all sessions of one IP address, no matter how many or how few, can be compressed into one unit representation. In addition, we also segmented sessions in the largest session-aggregation cluster to identify subpatterns of this cluster. Although we cannot rule out the possibility that one IP address was used by multiple users, this is arguably a less likely scenario to our best knowledge of CiteSpace usage. For example, the majority of users are researchers and students, who typically use their own laptops. This potential ambiguity can be resolved with additional information, for example, user registration information. This is one direction of our future work.

IP address mapping. In addition to session identification, we map logged IP addresses to geographical locations. Matching IP addresses to geographical locations is much more accurate at the country level than at the city level. At the city level, the accuracy typically ranges from 50%[8] to 88% (Guo et al., 2009). In contrast, at the country level the accuracy can reach ~96% to 98% (Poese, Uhlig, Kaafar, Donnet, & Gueye, 2011).

[8]http://geoipinfo.org/

DOI: 10.1002/asi

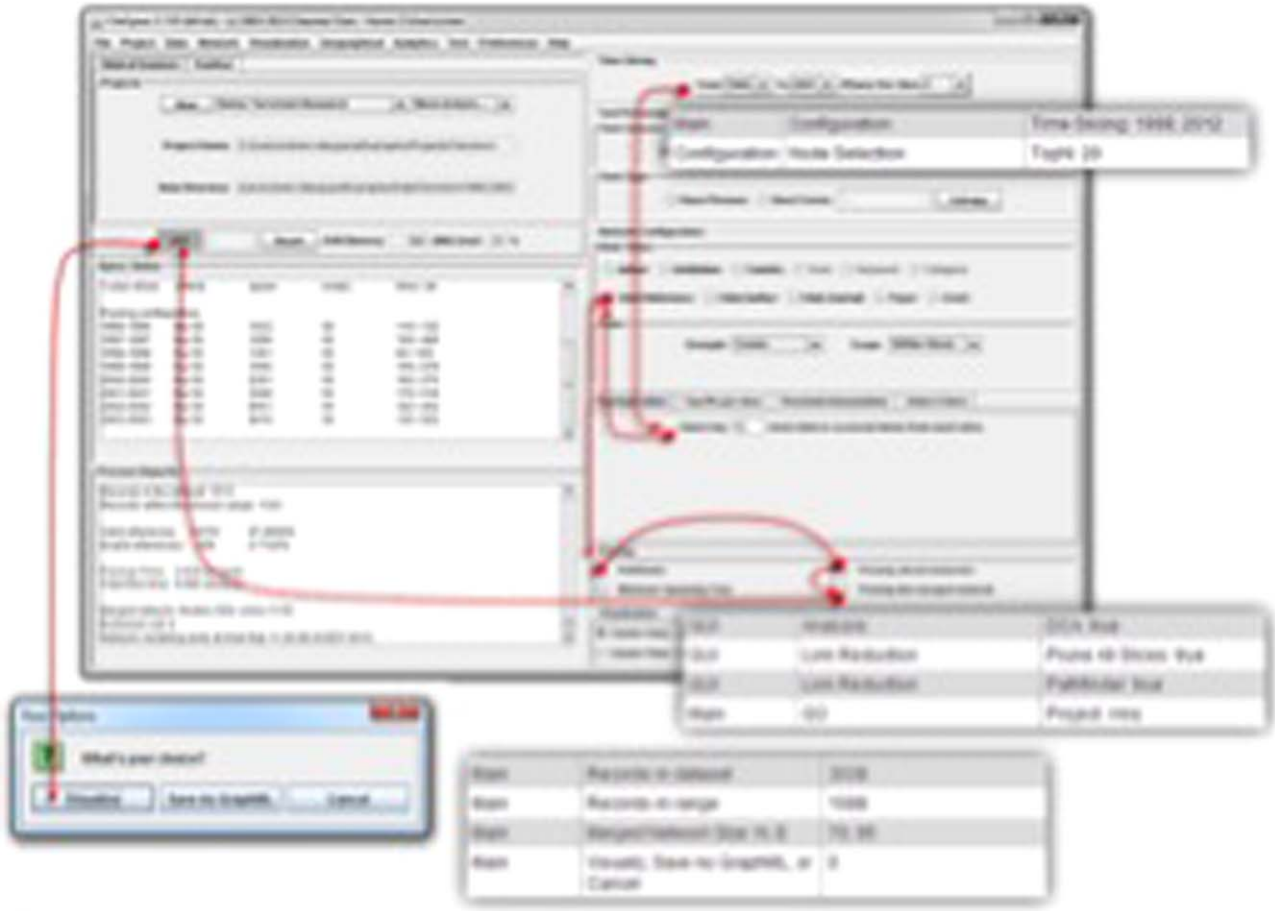

FIG. 1. A clickstream in the main user interface of CiteSpace and associated log events. [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2. Attributes of logged events in CiteSpace.

Attribute	Definition	Example
IP address	IP address of the user	1.222.333.444 (artificial IP address)
Time	Timestamp of the triggered event	2014-05-06 08:55:28
Context	Area on the GUI where the event is triggered	Main window; GUI window
Event	Name of the event	Configuration, node selection, link reduction
Value	Value of the event	TopN: 20

Clusters of IP-Aggregated Sessions

To cluster aggregated sessions, we must determine how to represent an IP-aggregated session, how to measure the similarity between different IP-aggregated sessions, and how to group these IP-aggregated sessions.

IP-aggregated session representation. We chose to represent each IP-aggregated session with a probability distribution of all 81 types of events during the entire data collection window. Each IP-aggregated session is characterized by a vector of the probability of each of the events. The vectors of all the IP-aggregated session together form an IP-aggregated session-event matrix $M_{session\ \ aggregation-event}$, as shown in (1):

$$M_{session\ \ aggregation-event} = \begin{bmatrix} p_{sg_1e_1} & \cdots & p_{sg_1e_n} \\ \vdots & \ddots & \vdots \\ p_{sg_me_1} & \cdots & p_{sg_me_n} \end{bmatrix} \quad (1)$$

where $M_{session\ \ aggregation-event}$ is the IP-aggregated session-event matrix with m IP-aggregated sessions by n events. The $p_{sg_ie_j}$ in the matrix represents the probability of IP-aggregated session i associated with event j.

DOI: 10.1002/asi

IP-aggregated session similarity. Given the probability distribution of the events of each IP-aggregated session, the Kullback–Leibler Divergence (KL Divergence), also known as the relative entropy, is used to measure the dissimilarity between two IP-aggregated sessions. The KL Divergence measures the nonsymmetric divergence of two probability distributions P and Q (Kullback & Leibler, 1951), and $D_{KL}(P \parallel Q)$ represents the information loss using Q to approximate P (Burnham & Anderson, 2002):

$$\mathrm{D_{KL}(P \parallel Q)} = \sum_{\mathrm{i}} \mathrm{P(i)} \ln \frac{\mathrm{P(i)}}{\mathrm{Q(i)}} \tag{2}$$

P and Q are the probability distributions of IP-aggregated session SG_P and SG_Q, respectively. We take the average of $D_{KL}(P \parallel Q)$ and $D_{KL}(Q \parallel P)$ as a symmetric difference between P and Q, which is also called the Jensen–Shannon Divergence or the Information Radius (Manning & Schütze, 1999):

$$d(P,Q) = \frac{D_{KL}(P \parallel Q) + D_{KL}(Q \parallel P)}{2} \tag{3}$$

Then, an IP-aggregated session dissimilarity matrix D is derived from (3) by calculating $d(sg_i, sg_j)$, the dissimilarity between sg_i and sg_j:

$$D = \begin{bmatrix} 0 & \cdots & d(sg_1, sg_m) \\ \vdots & \ddots & \vdots \\ d(sg_1, sg_m) & \cdots & 0 \end{bmatrix} \tag{4}$$

Clustering method. We apply an agglomerative hierarchical clustering algorithm to the IP-aggregated session dissimilarity matrix D, defined on the basis of KL Divergences. Agglomerative hierarchical clustering is a "bottom-up" approach to merge pairs of clusters recursively (Rokach & Maimon, 2005). Ward's method is used to choose the next pair of clusters to merge (Ward, 1963).

To determine the number of clusters, we use the Average Silhouette Width (ASW) score to evaluate the appropriateness of clustering at different numbers of clusters. The ASW score evaluates how well each item is clustered based on the ratio of its within-cluster distance and between-cluster distance. Given an item, an ASW close to 1 suggests a good clustering result, a value close to 0 suggests that it may belong to two clusters, and a value close to -1 suggests a wrong membership (Rousseeuw, 1987).

We visualize the resultant clusters using multidimensional scaling and color the members of these clusters accordingly.

Interactive Pattern Profiling

We define an interactive pattern in terms of a series of state transitions in the state space of 81 types of logged events. For example, how would one traverse the state space? What paths are deemed problematic? How would the paths of one type of IP-aggregated sessions differ from another? These are the questions we want to answer to profile each IP-aggregated session cluster.

Interactive pattern profiling is an exploratory process. Initially, we have no information about whether a cluster of IP-aggregated sessions represent a particular level of proficiency or a grouping of tasks. We derive characteristics of a pattern profile from multiple dimensions, including the amount of usage, the coverage of the state space, the reach of crucial events, and state transition patterns.

Amount of usage. The amount of usage of an IP-aggregated session is measured as the total number of occurrences of events triggered within this IP-aggregated session during the entire data collection period. It is perhaps the most fundamental feature of a session-aggregation profile. We assume that an IP-aggregated session with longer intervals and more interactive events would be associated with more sophisticated behavioral patterns, or more complex tasks.

Coverage of the state space. The coverage of the 81-event state space indicates the extent to which an IP-aggregated session has involved various functionalities of CiteSpace. The complexity of each of the 81 types of events varies from simple ones to those with relatively steep learning curves. We expect that a user behind a simple IP-aggregated session would probably visit fewer states of the space than a user who generated a sophisticated IP-aggregated session.

Reaching crucial events. Each visual analytic application is designed to support a range of tasks. Some tasks are crucial, whereas others may be peripheral. CiteSpace is specifically designed to help users make sense of the structure and dynamics of a scientific domain. Some tasks are essential to achieve this goal, such as applying clustering algorithms to a network and generating labels for resultant clusters. We note that IP-aggregated sessions may differ substantially in terms of whether they have reached certain areas in their state transition history. From an analytic point of view, the most telling event in CiteSpace would be that associated with labeling clusters using LLR, involving the use of the logarithm likelihood ratio test to select labels that can best represent the central theme of a cluster (Chen, 2006). We examine whether the occurrence of these crucial events differs significantly across different IP-aggregated session groups.

State transition patterns. A state transition pattern of an IP-aggregated session is a directed graph defined on the space of the 81 states based on all the transitions recorded from the sessions of interaction. For example, a hypothetical transition path could be *Configure→Visualize→Cluster Generation→Label Generation*. Given finite states (events), state transitions could be modeled as a first-order Markov Chain process, where the probability of each state depends

DOI: 10.1002/asi

only on its previous state (Norris, 1998). For each individual IP-aggregated session, a state transition graph could be considered the IP-aggregated session's distinct "behavioral signature." By aggregating all IP-aggregated sessions' state transition graphs in the same cluster, one can study typical state transition patterns associated with each IP-aggregated session group.

Specifically, we calculate an 81x81 Markov state transition matrix for each IP-aggregated session with each element being the probability of transition from event i to event j within each session. Then we compute the aggregated state transition matrix within a cluster C by averaging each element over all matrices within the cluster.

We visualize the aggregated state transition matrices of all the clusters as directed graphs so that we can compare the behavioral patterns of IP-aggregated sessions. Different behavioral patterns are also matched with different task progresses in CiteSpace.

Segmentation of the biggest cluster by session quartiles. We perform segmentation of all the individual sessions to discover subtle differences of behavioral patterns within the biggest cluster. First, we define the length of a session as the total number of state transitions, that is, how many jumps a user made within the session. Second, we ranked all the sessions by their lengths and divide them by quartiles. Third, we analyze the aggregated state transition patterns of the four quartiles to see if there are subtle differences between the four quartiles of sessions within this biggest cluster.

Version Adoption Pattern Analysis

The current log data contain the usage of 16 versions of CiteSpace. The majority of the differences among these versions are relatively small, with a few exceptions where substantial changes are involved. It is entirely up to the users to decide whether they should switch to a new version.

Will IP-aggregated sessions from different clusters have different patterns regarding this choice? We conducted a survival analysis to answer this question. For an individual IP-aggregated session, the duration of survival is defined as the duration between the current version number that appears in the IP-aggregated session's earliest session and the session in which the next version number appears for the first time.

Specifically, to derive the survival days for survival analysis, we subtract the start date from the end date for IP-aggregated sessions associated with each version of CiteSpace. Then the IP-aggregated session is assigned a state of "1" if it switched to a newer version. For the latest versions, such as 3.8.R6 (64-bit), an IP-aggregated session's state is censored as "0" because it is the most recent status that our log event data can observe. Then we plot the usage of different versions of the entire user population and select the survival function of the three most representative versions of different IP-aggregated session clusters. The median number of survival days for the three representative versions are estimated.

Identifying Error-Prone Areas

Error-prone areas are areas in the state space of CiteSpace where users tend to make mistakes. Finding error-prone areas can help trouble-shooting problems and improve the design. However, it is difficult to tell why some areas are error-prone based on usage logs alone. Therefore, we gathered additional information from a personal blog maintained for users learning CiteSpace.[9] The blog is in Chinese because a large number of users are from China. At the time of writing, the blog logged 424,787 visits and was followed by 634 users. Readers of the blog can post comments and ask questions about CiteSpace. We crawled 1,480 comments posted by users of CiteSpace from January 2011 to April 2015 on the CiteSpace blog.

First, in order to map the state transition patterns to the real-world problems reported by users, we introduce the notion of "transitional points." A "transitional point" is an event that connects a lower-proficiency pattern to a higher-proficiency one. Such transitional points stand out in both the usage logs and blog data. We cross-reference issues found in user feedback and user reports with user behavioral patterns associated with transitional points in order to explain what had most likely happened during those transitional points. For example, if the simplest behavioral pattern ends with "Go," did errors occur at those moments?

Second, in order to have an overall picture of user issues, we clustered these user comments using the Lingo algorithm (Osiński, Stefanowski, & Weiss, 2004) implemented with Carrot2 Java API.[10] This gave us an overview of the salient topics and subtopics frequently discussed by users in the blog. The initial clustering identified many broad topics. To obtain more details, we used an iterative clustering method controlled by a predefined threshold of the size of a cluster ($n \geq 40$).

Third, we summarized four high-level themes of challenges users encountered through content analysis. Each user comment (feedback) was labeled with one or more concepts, and concepts were merged into themes.

Results

In this section, first, we report various results regarding the usage, duration, and geographical distributions of users and the major topics that users investigate using the CiteSpace blog. Second, we summarize the IP-aggregated session clusters and the interactive and adaptation patterns associated with each cluster. Finally, we describe the error-prone areas in CiteSpace through content analysis of user feedbacks at critical transitional point events.

[9]http://blog.sciencenet.cn/u/ChaomeiChen
[10]http://project.carrot2.org

DOI: 10.1002/asi

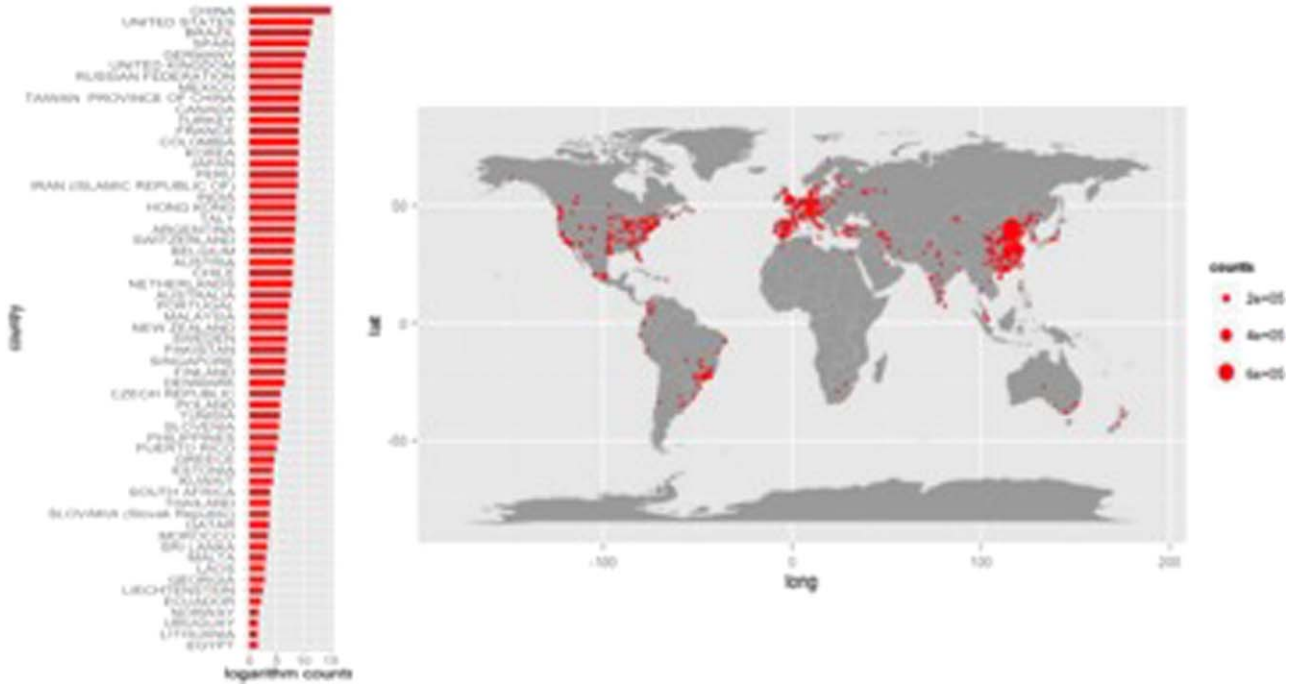

FIG. 2. Geographical distribution of logged events. [Color figure can be viewed at wileyonlinelibrary.com]

Primary Data Statistics

The duration of the entire data collection period was 439 days, from August 4, 2013 to October 17, 2014. The users' demographic information is very limited because the only information we have from the log data is IP addresses. We made an effort to map IP addresses to geographical locations using the API of a free IP-locating service.[11] We were able to map the 18,049 unique IP addresses to 792 cities in 59 countries (Figure 2). The size of a circle on the map represents the intensity of use in a city, that is, the total number of events aggregated in that city. The length of the red bar on the left represents the natural logarithm-transformed total number of events that occurred in that country. The country with the most intensive usage of CiteSpace is China. The cities with the most intensive use are major Chinese cities, such as Beijing and Shanghai. The second-highest usage is in the United States, particularly in the Northeast and along the West Coast.

As shown in Table 3 the number of events represents how many instances of events each IP address triggered during the entire data collection period and the duration is the number of days between the earliest date and the most recent date logged for each IP address. On average, an IP address triggered 225.26 events and spent 16.8 days using CiteSpace during the entire period of study.

The top 30 most popular topics analyzed in CiteSpace, ranked by the number of Label:LLR events, are shown in Table 4. The topic is derived from the label of the largest cluster in CiteSpace. The label of a cluster is extracted from articles that cite members of the cluster. The most extensively used topic is "Bioterrorism," which is the largest topic in the sample data set provided along with CiteSpace. Not surprisingly, this is the most commonly analyzed data set. The fourth most popular topic is "Subject Review" or "Review," which is in accordance with what CiteSpace is designed for—help users to understand a body of literature with a visual analytic tool. The list shows a diverse range of topics. Among them are three relatively long titles in Chinese, which reveal rich information about the topics to be explored.

Figure 3 shows the geographical distributions of IP addresses associated with some of the most popular topics. For example, "Social," "Safety Climate and Safety Behavior—The Mediating Role of Psychological Capital," and "Subject Review & Review," are scattered around the world in several countries. Interestingly, the same topic was analyzed by different users from different cities. While it is conceivable that the same user worked on CiteSpace while traveling extensively to all these countries, an alternative interpretation that multiple users were involved is also reasonable. The latter scenario could lead to opportunities for communication and collaboration among users who are likely interested in the same subject. Furthermore, given the diversity of users' levels of proficiency and domain expertise, the potential for users to collaborate across different levels of experience appears to be a promising avenue to serve the user community of CiteSpace.

IP-Aggregated Session Clusters

We determine the optimal number of clusters based on the average silhouette width (ASW). The ASWs for k

[11]http://ipinfo.io/

DOI: 10.1002/asi

clusters from 2 through 15 are plotted in Figure 4. The two-cluster solution has the highest ASW value, that is, $ASW_{k=2} = 0.58$, followed by the three-cluster solution $ASW_{k=3} = 0.42$. The size of the clusters is uneven when $k = 2$. The largest cluster has 13,954 IP-aggregated sessions, and the other cluster has 4,094 IP-aggregated sessions. IP-aggregated sessions vary widely, ranging from one or two clicks to hundreds of state transitions. To provide a more detailed characterization of IP-aggregated sessions, we decide to retain the agglomerative hierarchical tree at $k = 3$, which leads to three clusters: Cluster #1 (4,094), Cluster #2 (3,594), and Cluster #3 (10,410). We will refer to these clusters as Clusters #1, #2, and #3 throughout the article.

Next, we apply the agglomerative hierarchical clustering with Ward's method to the dissimilarity matrix of all the IP-aggregated sessions. Figure 5 shows a dendrogram of the hierarchical clustering. The three rectangles in red mark the three-cluster cut. The dendrogram also supports the selection of three clusters.

Figure 6a shows a multidimensional scaling (MDS) map of all IP-aggregated sessions. Each point represents an IP-aggregated session. Figure 6b shows the same MDS configuration colored by the membership of the three clusters, that is, Cluster #1 (red), Cluster #2 (green), or Cluster #3 (blue). The size of a point is proportional to the square root of the number of events associated with the IP-aggregated sessions. The MDS map and the hierarchical clustering results are consistent.

TABLE 3. Statistics on number of events and durations.

			95% Confidence interval for mean	
Variable	Mean	Std. error	Lower bound	Upper bound
Number of events	225.26	7.663	210.24	240.28
Duration (days)	16.81	0.414	16.00	17.62

Interactive Pattern Profile

We characterize IP-aggregated sessions in terms of the amount of usage, coverage of the state transition space, reach of crucial analytic events, and aggregated state transition patterns.

Usage. The amount of usage refers to the total number of events in an IP-aggregated session. Figure 7 shows the mean and standard deviation (SD) of the log-transformed number of events for each of the three clusters. Cluster #1 has the lowest usage (mean = 2.23, SD = 0.86), Cluster #2 is intermediate (mean = 3.88, SD = 0.88), and Cluster #3 has the highest usage overall (mean = 4.93, SD = 1.23).

Previous research suggests a correlation between expertise (in terms of the number of distinct events) and time spent on an application (Linton, Joy, & Schaefer, 1999). Therefore, we hypothesize that Cluster #1 represents sessions of low-proficiency, Cluster #2 represents sessions of intermediate-proficiency, and Cluster #3 represents sessions of high-proficiency.

Coverage of the state transition space. The state transition space has a total of 81 types of events. Table 5 shows the total number of event types associated with each cluster. Cluster #3 has the broadest coverage, 96.1% of the entire

TABLE 4. Top 30 most popular topics analyzed in CiteSpace.

Rank	Topic	# of Events	Rank	Topic	# of Events
1	Bioterrorism	4,452	16	ISI Web	174
2	Social	702	17	Ocular Injury	172
3	安全氛围对员工安全行为的影响———心理资本中介作用的实证研究 (Safety Climate and Safety Behavior—The Mediating Role of Psychological Capital)	579	18	生态文明 (Ecological Civilization)	167
4	Subject Review & Review	472	19	Road Transportation Companies	161
5	Information	404	20	Collaborative Logistics; Transportation Companies	160
6	Competitiveness	390	21	Science	160
7	Knowledge	332	22	Turbid Lake	160
8	Data	326	23	Lean Assessment Tool	160
9	Role	272	24	Personality	156
10	Innovation	238	25	Clustering	149
11	Tourism	234	26	Corporate Social Responsibility	142
12	Par2	224	27	教练员素质结构的研究现状与分析 (Survey on Analysis of Coach's Competence Structure)	140
13	Effect	219	28	Semantic	138
14	Application	217	29	Research	137
15	Analysis	205	30	Adolescents; Child Behavior Checklist; Teachers R	135

DOI: 10.1002/asi

FIG. 3. User distribution with different topics. [Color figure can be viewed at wileyonlinelibrary.com]

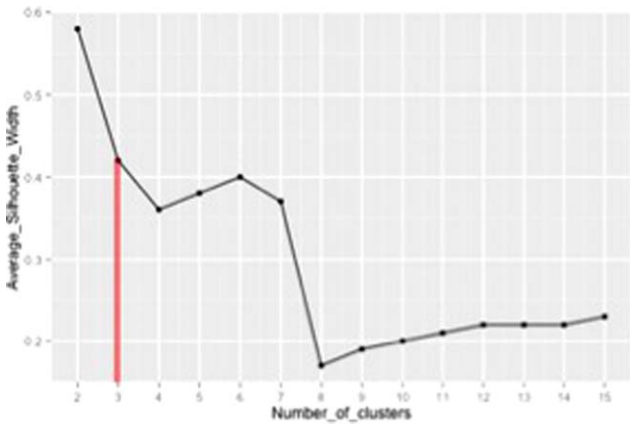

FIG. 4. ASW for the number of clusters (k = 2, . . ., 15) generated by hierarchical clustering. [Color figure can be viewed at wileyonlinelibrary.com]

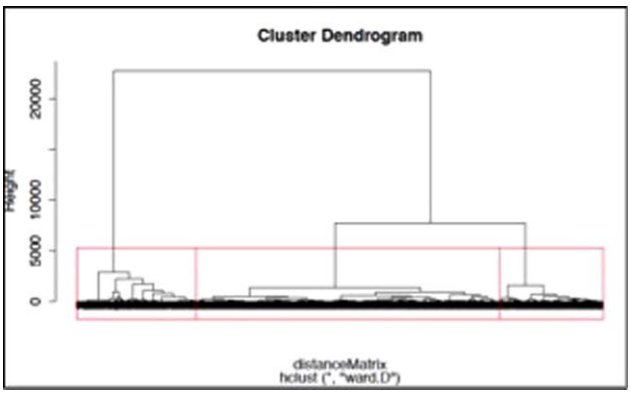

FIG. 5. The dendrogram showing the three-cluster cut on the agglomerative hierarchy. [Color figure can be viewed at wileyonlinelibrary.com]

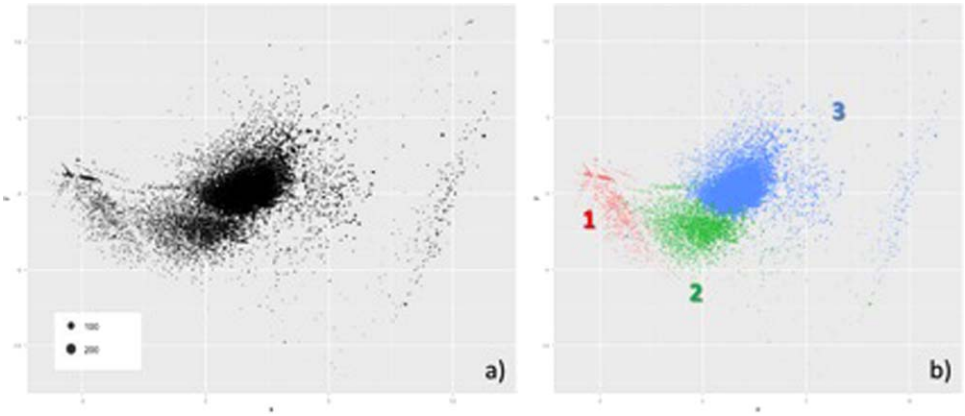

FIG. 6. MDS maps of IP-aggregated sessions (IP addresses): (a) the original MDS map, and (b) the MDS map colored by cluster membership. The size of a point is proportional to the square root of the number of events for an IP-aggregated session. [Color figure can be viewed at wileyonlinelibrary.com]

space, compared with Cluster #2 (61.7%) and Cluster #1 (87.7%).

Figure 8 shows how much each event was used in each cluster. Cluster #3 not only covers more types of events, but also, on average, triggers events more often (x-axis). In particular, the event "Interaction" is triggered much more often in Cluster #3 than in Cluster #1 and Cluster #2. This finding indicates that high-proficiency level IP-aggregated sessions considerably cover more functionalities than the other two clusters. Furthermore, high-proficiency level IP-aggregated sessions spend longer time interacting with CiteSpace.

In Figure 8, many events have relatively low frequencies. We will not discuss them in detail because our focus is on major state transition patterns. Nevertheless, there are several explanations of low-frequency events. For example, since the software has been actively maintained, newly

DOI: 10.1002/asi

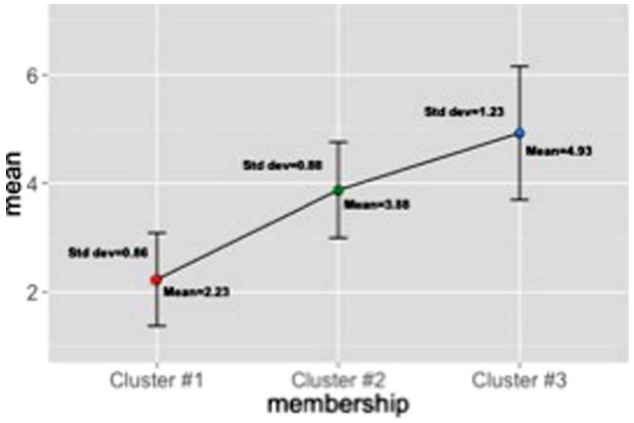

FIG. 7. Means and standard deviations of the logarithm of the numbers of events for the three clusters. [Color figure can be viewed at wileyonlinelibrary.com]

added features may not be used as frequently as features that have been available for a long time. In addition, some of the features are not fully documented in the user manual or tutorials. The awareness of such features tends to be relatively low. Furthermore, some tasks can be accomplished through alternative paths, which may reduce the perceived importance of some types of events along a particular path. Last but not least, previous research on the transition of user expertise suggests a "stabilization effect" among users. This effect refers to the phenomenon that after a long term of usage, even experienced users could be confined to only a limited set of commands of a system, probably because after learning even a subset of commands, they are able to perform their tasks well enough (Cockburn, Gutwin, Scarr, & Malacria, 2015).

Reaching crucial analytic events. Crucial analytic events are milestones of a visual analytic process. We focus on the occurrences of the event type "Label: LLR" in each cluster of sessions because an occurrence of the event is clear evidence that the user was able to access crucial information generated by CiteSpace concerning the structure and dynamics of a scientific domain. As shown in Table 6, 62.73% of sessions in Cluster #3 reached this level. In contrast, the rates are much lower for sessions in clusters of intermediate- (2.62%) and low-proficiency (less than 0.07%). This finding is consistent with our hypothesis in Results→Interactive Pattern Profile→Usage that IP-aggregated sessions in Cluster #1 are essentially at the low-proficiency level of a visual analytic process. IP-aggregated sessions in Cluster #2 make modest progress but are substantially behind those in Cluster #3.

State transition paths. A state transition graph represents a distinct "behavioral signature" of each IP-aggregated session. Figure 9 shows a combination of state transition graphs for the three clusters to highlight how sessions in the three clusters differ. Each node represents an event, and each arc with an arrow represents the direction of a state transition between two events. The graph is generated by averaging all the state transition probabilities over each cluster and retaining the average state transition probabilities above 0.1. The thickness of the arc represents the probability of state transition.

TABLE 5. Coverage of distinct types of events in three clusters.

Cluster	Number of types of events reached	Percentage of all types (81)
Cluster #1	71	87.7%
Cluster #2	50	61.7%
Cluster #3	77	96.1%

The state transition graph shows that the red path representing a typical Cluster #1 session starts with event "Version" but barely moves beyond "Projects" and then ends, perhaps prematurely, with the "Exit" event. The green path of a Cluster #2 session also starts with "Version" and reaches as far as the event "Visualize | Save As GraphML | Cancel," which indicates that users are able to configure CiteSpace to visualize a network. The blue path of a Cluster #3 session reaches even further, with a high probability to events such as "Control Panel," "Label: LLR," and "Interaction," indicating that users are able to analyze the structure and dynamics of a network. The graph reinforces our observation about how we may differentiate IP-aggregated sessions from the three clusters.

If we tie the different state transition patterns of the three clusters to actual task progress in CiteSpace, we can get the following task progress descriptions:

Cluster #1 (IP-aggregated sessions of low-proficiency)—Task Initiation. The task reflected in this state transition pattern is to initiate CiteSpace and to build up a project. Concretely, the task is to: 1) initiate the CiteSpace application with the events "Version" and "Build Date"; 2) consent to the agreement of CiteSpace with event "Consent"; 3) build up the project with the event "Projects." The "Link Reduction" is an automatic event that records the current default values of option for link reduction.

Cluster #2 (IP-aggregated sessions of intermediate-proficiency)—Visualization Generation. In addition to task initiation, the task reflected in this state transition pattern is to preconfigure parameters of the visualization and generate a visualization. More specifically, the task is to: i) select the term source and term type with event "Term Selection" and specify what type of network to be visualized with event "Analysis," for example, author cocitation analysis network or document cocitation analysis network; ii) start the automatic running process to generate the visualization, beginning with event "Go," followed by event "Configuration" to record the values of configuration and event "node selection" to record the value of selected node types, and then followed by the calculation of values of "Records in range" and "Records in data set," and ending with event "Merge network size: N;E" and event "Visualize | Save as GraphML | Cancel" to generate a visualization.

DOI: 10.1002/asi

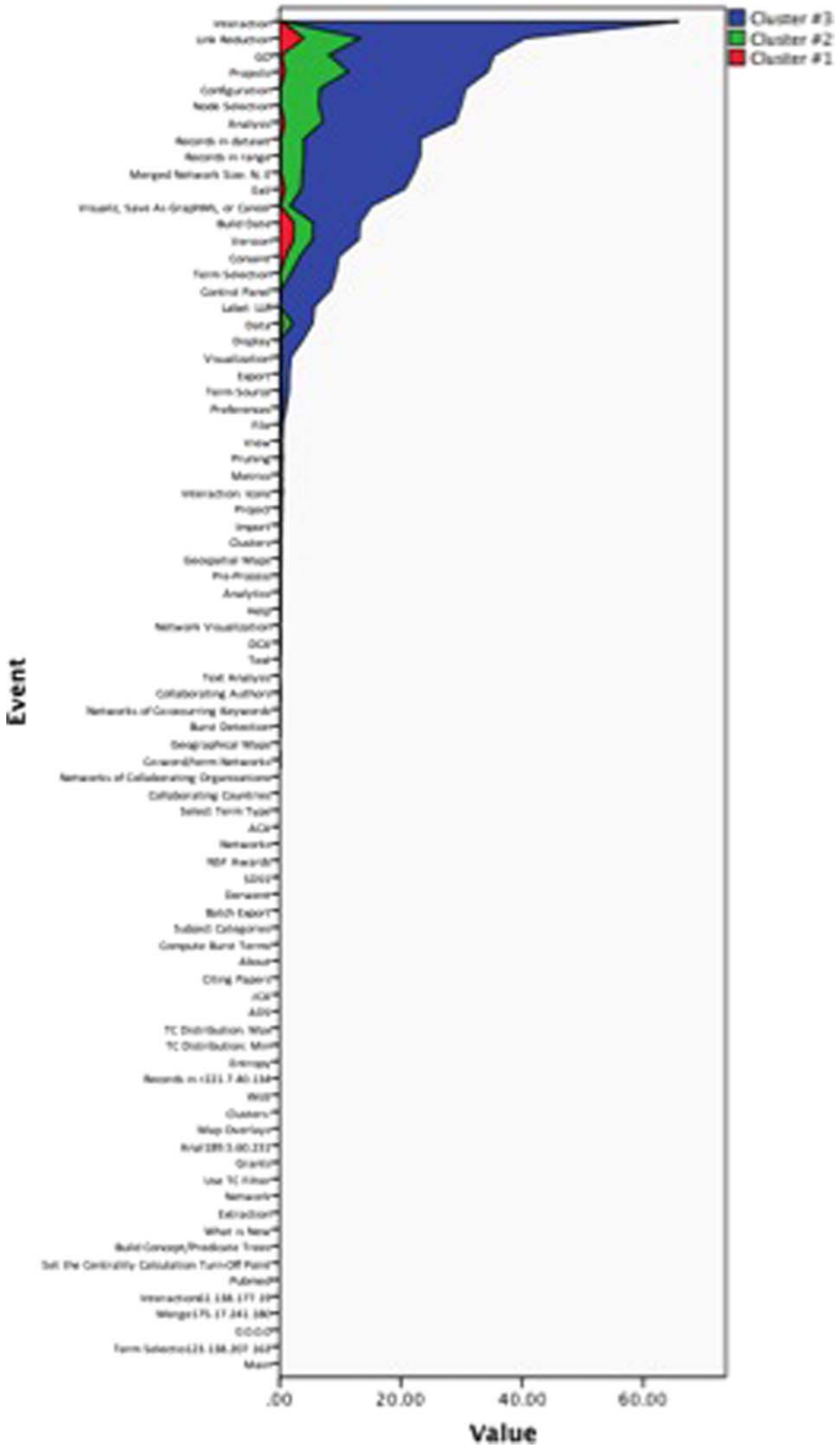

FIG. 8. Frequency distributions of event types in the three clusters (red: low proficiency, green: intermediate proficiency, blue: high proficiency). [Color figure can be viewed at wileyonlinelibrary.com]

Cluster #3 (IP-aggregated sessions of high-proficiency)—Visualization Adjustment and Labeling. The visualization generated in Cluster #2 is a network, the task reflected in Cluster #3 is to partition the network, perform adjustment operations on the network, and label the clusters. The task is to: 1) change how node labels are displayed by a

DOI: 10.1002/asi

TABLE 6. Percentage of IP-aggregated sessions Using "Label:LLR" in three clusters.

Cluster	Number of IP-aggregated sessions using LLR	Number of IP-aggregated sessions	Percentage of IP-aggregated sessions Using LLR
#1	3	4,094	0.07%
#2	93	3,544	2.62%
#3	6,530	10,410	62.73%

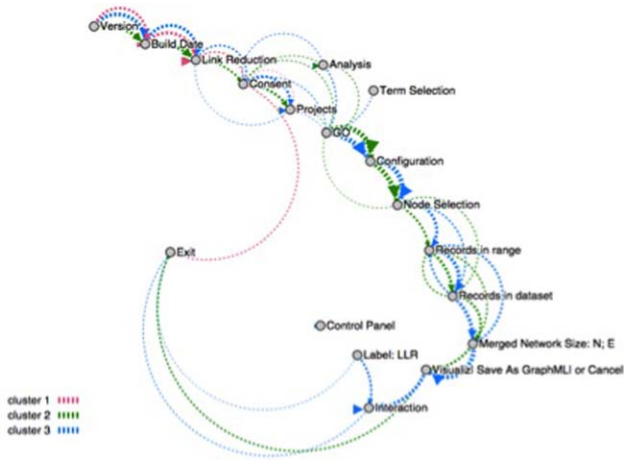

FIG. 9. Combined state transition graphs of the three clusters. [Color figure can be viewed at wileyonlinelibrary.com]

combination of threshold values and choose graph layouts with event "Control Panel"; 2) cluster the network and label the clusters with event "Label: LLR"; 3) adjust specific node in the network with options in event "Interaction."

Further analysis of the biggest cluster by session quartiles. Cluster #3 is the biggest cluster, with 10,410 IP-aggregated sessions, almost twice as large as cluster #1 (4,094) and cluster #2 (3,594).

First, we rank all the sessions within Cluster #3 by their length as a chain of state transitions. The histogram of session lengths is depicted in Figure 10. The distribution of session lengths follows approximately a power-law function. This means that a few sessions have triggered a great number of events while the majority (long tail) sessions have triggered much less.

Second, as shown in Table 7, if we segment all the sessions to four quartiles, 75% of sessions are relatively short; the length of a session tends to be fewer than 26 transitions.

Third, we aggregated and visualized the state transition path (threshold = 0.2) of sessions within each quartile range in Figure 11. The sessions in the first quartile contain a variety of two-step state transitions and feature one state transition of "Version"→"Build Date." The sessions in the second quartile contains as long as eight-step state transitions and feature the path from "Go" to "Interaction." The

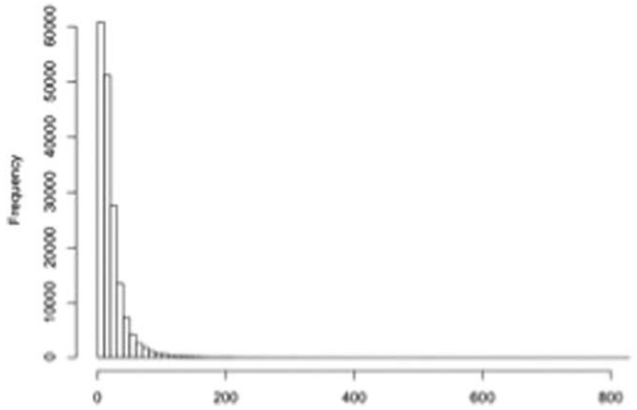

FIG. 10. Histogram of session lengths within Cluster #3 (X-axis: length of session; Y-axis: frequency of length).

TABLE 7. Values of four quartiles of the sessions within Cluster #3.

Quartiles	25%	50%	75%	100%
Lengths	2	8	26	824

sessions in the third quartile assemble the paths in the first and second quartile and add the state transitions of "Consent"→"Projects" and self-transition of "Analysis." The sessions in the fourth quartile assemble all the state transition paths in the first three quartiles and add one more state transition of "Lable:LLR"→"Interaction."

It appears that the two-step state transition pattern in the first quartile resembles the pattern of Cluster #1, and the state transition patterns in the second and third quartile resemble the pattern of Cluster #2 (with extra event "Interaction"). However, a further analysis reveals that while the most-frequent pattern of the first quartile of Cluster #3 is similar to Cluster #1, Cluster #3 includes less frequent patterns that are unique to Cluster #3. Sessions in the first quartile also explore many two-step transitions along the path from "Go" to "Interaction," although not very frequent, while sessions in Cluster #1 are unlikely to go beyond "Consent"→"Projects." Transition patterns in the second and third quartiles of Cluster #3 and Cluster #2 differ similarly.

How Different Versions Are Adopted?

An overview of the usage of different versions in the entire user population is shown in Figure 12. The x-axis represents different versions in the chronological order of the dates of their release. The y-axis represents the total number of IP-aggregated sessions ever using a particular version within the event log window. Older versions, such as 3.7.R1–3.7.R4, generally had many fewer IP-aggregated sessions than later versions, such as 3.7.R7–3.8.R1. The latest versions, such as 3.8.R2–3.8.R6, also have a smaller number of IP-aggregated sessions, probably due to the fact that they were only released recently. The 3.7.R8 (64-bit) and 3.8.R1 (32-bit) versions are by far the most popular ones.

DOI: 10.1002/asi

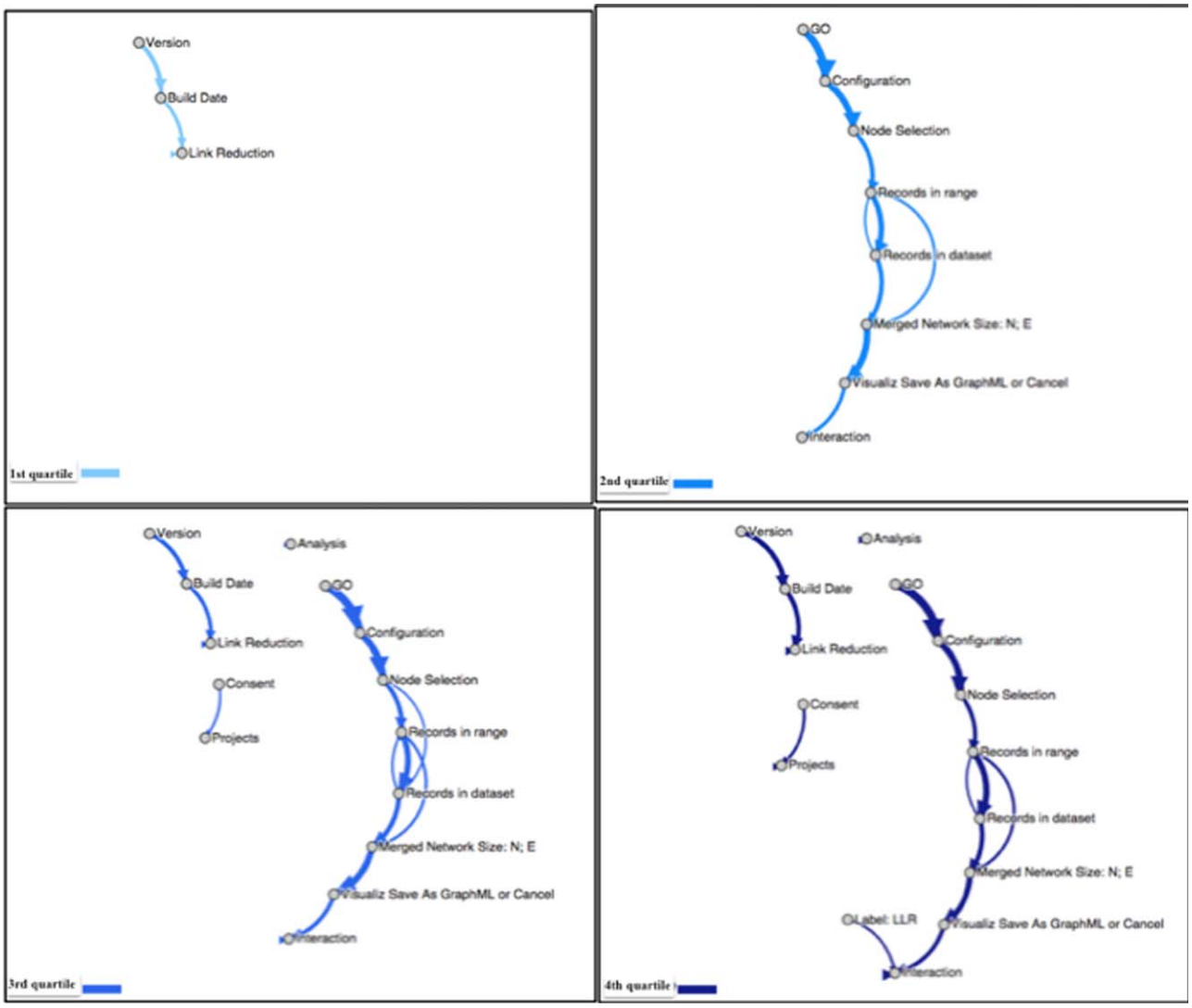

FIG. 11. State transition maps of the four quartile ranges within Cluster #3. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 13 shows three survival functions of the three most popular versions: left: 3.7.R7 (32-bit) with 3,211 IP-aggregated sessions, middle: 3.7.R8 (64-bit) with 5,707 IP-aggregated sessions, and right: 3.8.R1 (32-bit) with 4,396 IP-aggregated sessions. The blue lines are always the highest, the red lines the lowest, and the green lines in the middle. These patterns suggest that IP-aggregated sessions of high-proficiency in Cluster #3 (blue line) are more likely to stay with the current version than IP-aggregated sessions of intermediate-proficiency, which in turn are more likely to stay than the IP-aggregated sessions of low-proficiency. One plausible explanation is that the more engaged users are more likely to continue to use the same version until they complete their work at hand before they upgrade to a newer version, whereas users of the other two types of sessions would be more flexible to switch to newer versions.

The survival test shows that the survival functions of the three clusters are significantly different, with Log Rank $\chi^2(2)=13.476$, Sig.$=0.001$ for 3.7.R7 (32-bit), Log Rank $\chi^2(2)=33.848$, Sig.$=0.00$ for 3.7.R8 (64-bit), and Log Rank $\chi^2(2)=44.606$, Sig.$=0.00$ for 3.8.R1 (32-bit).

Table 8 summarizes the median days of survival for the three most popular versions by cluster. For all three versions, users in Cluster #3 would typically use the same version twice as long as those in Cluster #2 and Cluster #1 before switching to new versions. Nevertheless, this finding by no means suggests that users exhibiting high-proficiency behavioral patterns are late adopters of new versions. In fact, in all three clusters the first adopter of a newly released version usually appears within the first 2 days of release.

Error-Prone Areas

First, we introduce the notion of "stage-transition" events. A "stage-transition" event in this context connects events of one stage to events of another. For example, in Figure 9 we identified two stage-transition events:

DOI: 10.1002/asi

- Event "Go" between "Task Initiation" and "Visualization Generation";
- Event "Visualize | Save as GraphML or Cancel" (referred as "Visualize" hereafter) between "Visualization Generation" and "Visualization Adjustment and Labeling."

Table 9 contains three categories of user narratives (translated from Chinese) discussing problems they encountered at each stage transition. For the first stage transition event "Go," we identified issues in the configuration stage. For the second stage transition event "Visualize," we found problems concerning interactions with the visualization. Even after the successful generation of visualization, users may still encounter problems (Table 9).

We clustered the user comments to obtain a bigger picture of the salient topics. Figure 14 shows a hierarchy of topics identified from the most commonly asked questions.

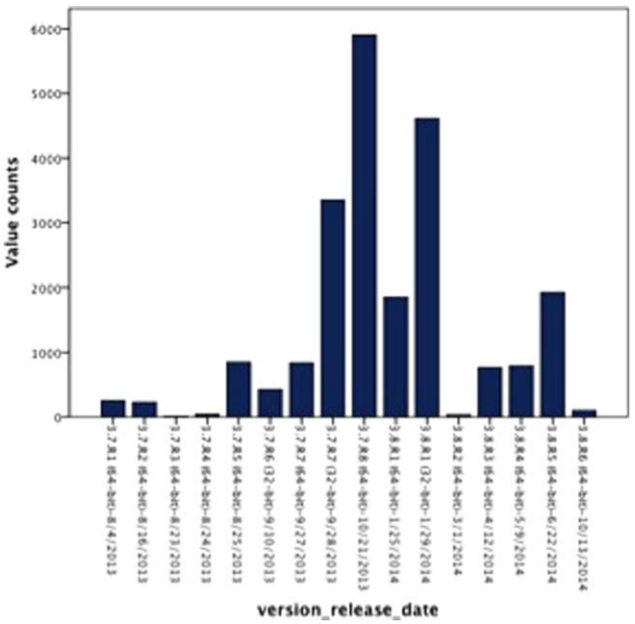

FIG. 12. Distribution of number of IP-aggregated sessions of each version. [Color figure can be viewed at wileyonlinelibrary.com]

Each node represents a cluster of comments, including questions. The size of a node indicates the number of comments in the corresponding cluster. Clusters with fewer comments are collapsed into a node labeled with "…". Clusters at higher levels represent broader areas than clusters at lower but more specific levels. The "downloading" cluster, for example, is concerned with downloading files from the Web of Science, CNKI, and PubMed. Some of the labels appear repeatedly in different contexts. For instance, the label "CSSCI" appears in topics such as "Author," "File," and "Transforming Data," indicating that users have questions concerning these aspects of the same data source CSSCI. These areas are valuable for improving the current design of CiteSpace further.

We summarized four high-level themes of challenges users encountered through content analysis. (Table 10). The first theme refers to confusions of similar concepts, particularly when users must make a choice from a few options associated with similar concepts. This theme underlines the role of knowledge of both bibliometrics and the visual analytic system; however, users may lack this type of knowledge. The second theme refers to the inconsistencies between the results produced by the system and users' expectations. For example, some users were confused about the differences between local and global citation counts. This inconsistency reflects the gap between users' mental model and the actual system model. The third theme refers to the need for more detailed instructions on certain operations. This need particularly indicates a lack of knowledge of the system. The last theme is the errors reported by users, including selecting a wrong operation, using incompatible versions, or inappropriate data formats. Error messages are particularly valuable for users to identify the error and learn how to correct it.

Discussion

This section discusses the interpretations and implications of the results with reference to the four questions raised at the beginning of the article.

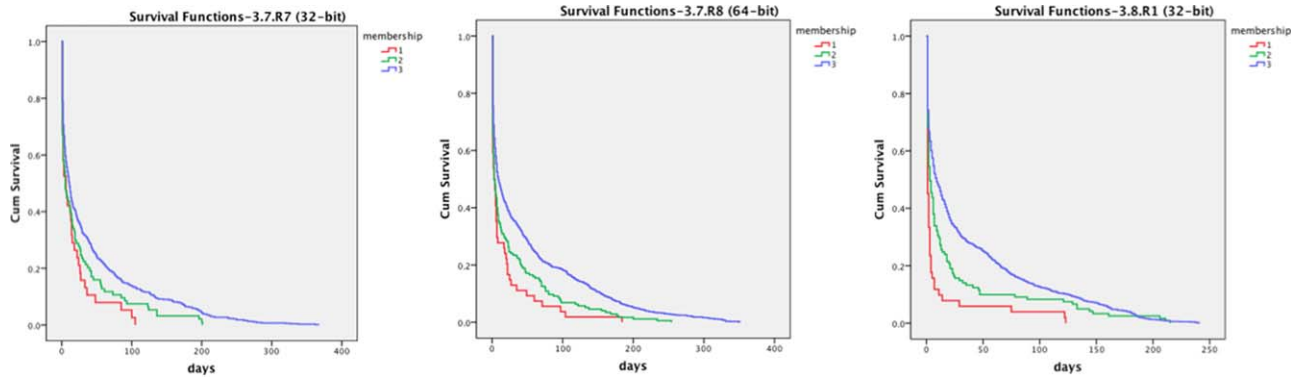

FIG. 13. Survival function plot for three representative versions of the three clusters. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

To What Extent Can We Discover User Behavioral Patterns From Usage Logs of CiteSpace?

We were able to discover three groups of major behavioral patterns and four subgroups within the most sophisticated sessions from the collected usage logs of CiteSpace, using our multimethod approach.

The three-cluster division evidently characterizes user behavioral patterns reasonably well in terms of the hierarchical clustering dendrogram (Figure 3) and the homogeneity of the groupings on the MDS map (Figure 4). Each cluster is associated with a distinct behavioral pattern and a set of profile facets, such as number of events, coverage of event types, reach of crucial events, as well as adaptation patterns. The profiles of different clusters are not only distinct from each other but also very consistent within clusters, indicating that we are revealing genuine properties of user behavior.

The four-subgroup division helped us to understand subtle behavioral differences within the largest cluster of sessions at the high-proficiency level. We observed that the high-proficiency level state transitions are not always perfect and complete, and they may contain side tracks, exploratory paths, and trials and mistakes, which all stand out from the sessions of low and intermediate proficiency because they have reached advanced events never reached by the first two groups of sessions.

TABLE 8. Coverage of distinct types of events in three clusters.

Version	Cluster membership	Median	
		Estimate	Std. error
3.7.R7 (32-bit)	#1	6.00	3.078
	#2	5.00	2.201
	#3	11.00	1.046
3.7.R8 (64-bit)	#1	3.00	1.470
	#2	3.00	0.863
	#3	9.00	1.111
3.8.R1 (32-bit)	#1	1.00	0.000
	#2	3.00	0.733
	#3	9.00	0.957

What Are the Characteristics of Each Group's Interaction Patterns?

The concept of interactive pattern is abstract. Under this concept, we explored multiple aspects or facets that we consider interaction patterns. The revelation of these facets is important in two ways: we constructed a multifaceted profile for each session cluster, and we confirmed that clusters are meaningful with correlated and consistent patterns on different facets.

We were able to construct a behavioral pattern profile that led to the identification of distinct interactive patterns in each cluster even without any prior knowledge of what each cluster represents. For example, we now have a much better understanding of high-proficiency level behavioral patterns: it is likely to invoke a wide range of events frequently (138 on average), have a good chance of arriving at crucial events (63% for Label: LLR), and routinely reach analytically

TABLE 9. User narratives and related issues regarding marking events.

Event	User Narratives	Issues
"Go"	*"...After I installed CiteSpace, I always got the error message when importing data after clicking 'Go': 'no data files found. check the following:(1) the data directory in the current project is correctly specified (2) data files must be named as download*.txt, e.g.download_2008a.txt' ..."*	Project Configuration
"Visualize"	*"...When I clicked 'Go', the application had been running for a while, but then the application froze. After I clicked 'Stop' I found that the JVM memory used was 99%. How could I adjust the configuration? Where can I change the value of JVM memory? (Note: my operating system is 32-bit windows XP, with 2G RAM. CiteSpace version: 3.8.R1 (32-bit) JRE Version: 1.7.0_67-b01)..."*	JVM Memory
	"...the application was frozen after I clicked 'Go'. I've tried this data set on multiple desktops, and it always froze. Then I tried a small proportion of the data set, and the application ran very well..."	Configuration
	"...However, after I clicked 'Go', the application didn't produce any results, and the space status report showed all zeros. The warning message was: 'make sure that your data files indeed include relevant information, such as references; try again with lower thresholds'..."	Missing Necessary Information
Post-"Visualize"	*"...After the visualization, the graph was constantly changing. When should I stop the process? Does it make any difference to the clusters discovered later on?..."*	Interaction
	"...How can I view information of all nodes after the visualization is generated? I tried multiple actions, but they are not what I want..."	Interaction

DOI: 10.1002/asi

important states such as "Interaction" and "Label:LLR" in state transition paths.

The discovered profiles further help us to determine whether the clusters are meaningful. If the clusters exhibit inconsistent or even conflicting patterns, for example, with sessions triggering the greatest number of events but covering only a small area of the state transition space or never reaching any crucial events, then it would be difficult to reconcile the conflicts and determine what each cluster actually represents. Notably, in our case, almost all the facets of the profile are so consistent with each other that we are comfortable giving a name to each cluster based on the existing evidence.

The fact that different levels of proficiency or behavioral patterns exist in CiteSpace is consistent with our understanding that CiteSpace is a function-rich visual analytic system with a relatively complex state transition space. High-functionality systems (HFS) are professional systems that are complex and perhaps challenging to learn at first, but can be learned to perform a wide range of tasks over time (Fischer, 2001). In such systems, how well or smoothly a user interacts with the system depends on the overlap between a user's mental model (knowledge about the domain and the system) and the system model. However, it is common that visual analytic applications can suffer from the mismatch between a user's mental model and system model. For example, users were sometimes confused by the naming of functionality in Prefuse (Heer et al., 2005), or felt lost in the complex interface of Action Science Explorer, or lacked sufficient knowledge of the domain to start the analysis in Action Science Explorer (Gove et al., 2011). One of the earlier usability studies of CiteSpace showed that users with a higher level of domain knowledge spent more time interacting with CiteSpace, and perceived CiteSpace as more useful (Yuan et al., 2013). Since CiteSpace demands user's knowledge of both the domain to be analyzed and that of how to make use of the various functionalities provided by the system, it is not entirely a surprise that groupings of multiple distinct state transition patterns emerge from the log analysis.

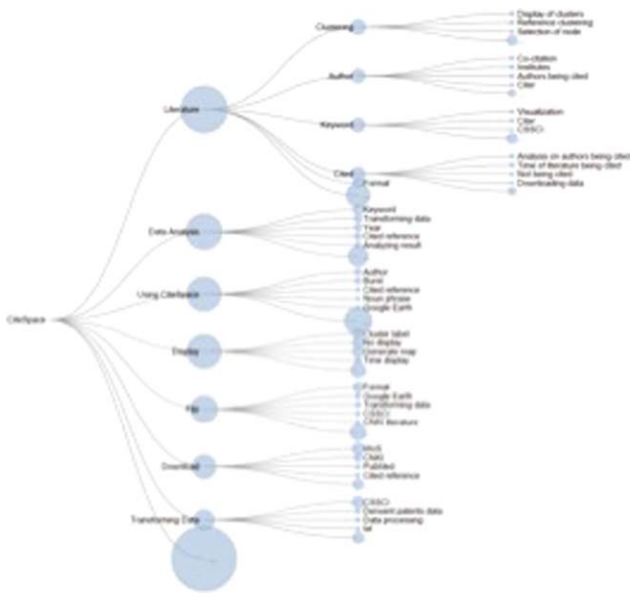

FIG. 14. The visualization of clusters of user comments. [Color figure can be viewed at wileyonlinelibrary.com]

What Are the Characteristics of Each Group's Adaptation Patterns?

The patterns of adapting new versions may appear to be counterintuitive initially but have reasonable interpretations. Users at the high-proficiency level took much longer to switch to a newer version than users at the entry and intermediate levels, at least for the three most intensively used versions. A plausible interpretation is that users at the high-proficiency level are more likely to be engaged in actual visual analytic processes, unlike users at the other two levels, who are essentially still in an early stage of learning. The hidden cost of switching would therefore be higher for users in the middle of an analytic project than for users who are merely learning the basics of the system.

Open-source software is usually maintained and updated for a diverse community of users with different levels of skills and experiences. CiteSpace, although not open-sourced, shares a great deal of similarity with open-source software because it is distributed freely. The present study reveals some of the intriguing but latent patterns that can help us to better understand the acceptance and adaptation of open-source or freely distributed software as it is being used by its diverse user community to meet their own practical needs. Previous research has identified individual skills and motivation as important factors in adoption of

TABLE 10. General themes from user questions.

	Theme	Paraphrased example
1	Users are confused by definitions of concepts, particularly among a group of similar concepts.	What are the differences among burst terms, cluster labels and noun phrases?
2	Users find results that are not consistent with their expectations.	The value of "record in the data set" is more than expected.
3	Users need detailed instructions to perform certain operations.	How can I find the modularity value? How can I find details of cluster members?
4	Users encounter errors during operations and need solutions.	Making errors in data import, selecting noun phrases and generating burst terms.

DOI: 10.1002/asi

innovations on the users' part (Greenhalgh, Robert, Macfarlane, Bate, & Kyriakidou, 2004), and software quality, systems capability, and software flexibility as critical factors in perceived usefulness and ease of use on the system part (Gallego, Luna, & Bueno, 2008). Our study provides some concrete evidence that advanced users of a visual analytic system may choose to delay their adoption of new versions, especially when existing versions adequately support their tasks. It would be interesting to explore whether different levels of proficiency correspond to different levels of individual skills and expertise, and how and why they are in turn associated with different rates of adoption. Furthermore, the technology adoption here is more of a "repeated" adoption of the updated software. This might help to explain the seemingly counterintuitive result that sophisticated behavioral patterns are associated with a slow adoption rate.

What Are the Error-Prone Areas of the Process of Using CiteSpace and Their Design Implications?

From the perspective of a successful workflow in CiteSpace, the error-prone areas usually happen at the stage transition events between levels of task proficiency. For example, the problems that prevent a user from passing through the "Go" event, which invokes the actual visualization process, are often due to issues in the configuration stage. Problems surrounding the "Visualize" event are likely caused by configuring the network models and selecting the right types of entities. Furthermore, after passing the "Visualize" event, users may still encounter problems when they interact with the visualization.

One interesting contrast is that the present study has revealed some types of errors that had not been identified in previous usability studies of CiteSpace. In previous studies (Allendoerfer et al., 2005; Yuan et al., 2013), three error-prone areas were identified, that is, missing the display of titles, the inability to display terms, and the unawareness one can move nodes around. Interestingly, these error-prone areas are all at the "Post-Visualize" stage. In another previous study (Synnestvedt & Chen, 2005), errors happened mostly at the "Post-Visualization" stage and concentrated on controls and interactions. These two studies were conducted in laboratory settings. Participants followed step-by-step instructions to complete specific tasks with controlled data sets. Therefore, users' activities were limited to clearly defined paths with well-prepared data sets. Thus, the chance that anything could go wrong was low, especially at earlier stages prior to "Go" and "Visualize." The present study, on the other hand, has removed all these constraints on how users may proceed with the visual analytic system. Users were free to analyze data sets from their own sources such as CNKI and CSSCI in Chinese. However, new-found data sources may not provide information that is necessary for running a standard analysis that CiteSpace has been optimized for long-established sources such as the Web of Science. The log analysis has the advantage of identifying this type of error at a large scale of use in real-world settings.

The design implications derived from the present study are multifold. First, the communication between a user and the system needs to be enhanced with more diagnostic analyses. For example, it is important for the system to detect whether a data set is appropriate (in compatible formats and a feasible amount) and offer specific remedies. Second, some users may benefit from additional guidance towards understanding the visual metaphor of an intellectual space and network topology. Third, the concept of some bibliometric terms can be better conveyed through concrete examples.

Finally, this study has some limitations and can be improved upon in the future. First, the present study takes IP-aggregated session as a proxy of individual users. The uncertainty in the IP-aggregated sessions can be reduced by soliciting extra information from users, such as requesting users to register for each session. Second, we focused on frequently used types of events and state transitions in this study, but not on less frequently occurring events. Third, our representation of an IP-aggregated session is a vector of events with probabilities of the events. An alternative but computationally more expensive representation could use a large and sparse matrix of state transitions for each IP-aggregated session. Community detection methods, such as edge-removal algorithm based on edge-betweenness (Girvan & Newman, 2002), modularity maximization (Newman, 2004), and stochastic block models (Karrer & Newman, 2011) could be employed.

Conclusion

Visual analytic systems may involve a steep learning curve because they require knowledge not only about the analytic domain but also about the design of a system. This learning curve has resulted in diverse behavioral patterns with various levels of proficiency. However, how users interact and adapt with the system has not been investigated across a diverse population over a long period of time. Existing research concentrates on the in-laboratory usability study of user behaviors, which usually spans a short period of time with a small population, but nevertheless carries higher costs and is sometimes perceived as intrusive. Other naturalistic methodologies collect data from users but exploit only the surface of the potential of data-driven user analysis.

The present study utilized usage logs to understand users' interactive and evolving patterns with CiteSpace. The usage logs were collected cumulatively for a much longer time (14.6 months) and from a much larger population (18,048 IP addresses) than in past studies and in a naturalistic manner (back-end logs). We first discovered three clusters of distinct behavioral patterns. Within the largest cluster, we further distinguished use sessions into sub-behavioral patterns with subtle differences through segmenting the sessions by quartiles. Then we analyzed the interactive pattern of the three clusters, for example, the amount of usage, the coverage of the types of events, the reach of crucial events, and the state transition pattern. With the multimethod triangulation

DOI: 10.1002/asi

approach, each cluster exhibited a distinct profile regarding these patterns. These distinct profiles enabled us to name the three clusters pattern of low, intermediate, and high proficiency. The transitional points between different levels of behavioral patterns were identified and tied to real problems during events reported by CiteSpace users via the author's blog. Moreover, we analyzed the evolving patterns of the three clusters, that is, the patterns of adaptation to new versions. We discovered that users with high-proficiency level behavioral patterns usually stick to their current versions for a much longer time than other users.

The contributions of the study are twofold. First, it provides a methodological demonstration of log analysis of usage for a visual analytic application through a multimethod approach. It triangulates behavioral patterns with multiple user characteristics and ties transitional points in behavioral patterns to real-world problems reported by users. The methodology used in the present study could be applied to other visual analytic applications, with minor modifications. Second, the interactive and evolving patterns discovered in the present study have built the foundation for a tool to improve CiteSpace in general. A user analysis tool that extracts user behavioral patterns and multiple behavior indicators from usage logs could promote identification of weakness and facilitate improvements of the software. The ultimate goal is to build an adaptive and intelligent system that can assist users in visual analytic tasks with a minimal cognitive burden.

Acknowledgment

This work is in part supported by the NSF I/UCRC Center for Visual Decision and Informatics (NSF IIP-1160960).

References

Allendoerfer, K., Aluker, S., Panjwani, G., Proctor, J., Sturtz, D., Vukovic, M., & Chen, C. (2005). *Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization*. In IEEE Symposium on Information Visualization (INFOVIS 2005). Washington, DC: IEEE Computer Society.

Arhippainen, L., & Tähti, M. (2003). *Empirical evaluation of user experience in two adaptive mobile application prototypes*. In Proceedings of the Second International Conference on Mobile and Ubiquitous Multimedia. New York: ACM.

Banerjee, A., & Ghosh, J. (2001). *Clickstream clustering using weighted longest common subsequences*. In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining. Society for Industrial and Applied Mathematics

Bavoil, L., Callahan, S.P., Crossno, P.J., Freire, J., Scheidegger, C.E., Silva, C.T., & Vo, H.T. (2005). *Vistrails: Enabling interactive multiple-view visualizations*. In Visualization, 2005. VIS 05. IEEE. Washington, DC: IEEE Computer Society.

Bayir, M.A., Toroslu, I.H., Cosar, A., & Fidan, G. (2009). *Smart miner: A new framework for mining large scale web usage data*. In Proceedings of the 18th International Conference on World Wide Web. New York: ACM.

Bostock, M., & Heer, J. (2009). Protovis: A graphical toolkit for visualization. IEEE Transactions on Visualization and Computer Graphics, 15, 1121–1128.

Burnham, K.P., & Anderson, D.R. (2002). Model selection and multimodel inference: A practical information-theoretic approach. Springer Science & Business Media.

Callahan, S.P., Freire, J., Santos, E., Scheidegger, C.E., Silva, C.T., & Vo, H.T. (2006). *VisTrails: visualization meets data management*. In Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. New York: ACM.

Cao, H., Jiang, D., Pei, J., Chen, E., & Li, H. (2009). *Towards context-aware search by learning a very large variable length hidden markov model from search logs*. In Proceedings of the 18th International Conference on World Wide Web. New York: ACM.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. Proceedings of the National Academy of Sciences of the United States of America, 101, 5303–5310.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for information Science and Technology, 57, 359–377.

Chen, L., Bhowmick, S.S., & Nejdl, W. (2009). COWES: Web user clustering based on evolutionary web sessions. Data & Knowledge Engineering, 68, 867–885.

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E. & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the American Society for information Science and Technology, 62, 1382–1402.

Cockburn, A., Gutwin, C., Scarr, J., & Malacria, S. (2015). Supporting novice to expert transitions in user interfaces. ACM Computing Surveys (CSUR), 47, 31.

Cook, K., Grinstein, G., & Whiting, M. (2014). The VAST Challenge: History, scope, and outcomes: an introduction to the Special Issue. Information Visualization, 13, 301–312.

Dimopoulos, C., Makris, C., Panagis, Y., Theodoridis, E., & Tsakalidis, A. (2010). A web page usage prediction scheme using sequence indexing and clustering techniques. Data & Knowledge Engineering, 69, 371–382.

Dogan, R.I., Murray, G.C., Névéol, A., & Lu, Z. (2009). Understanding PubMed® user search behavior through log analysis. Database, 2009, bap018.

Eccles, R., Kapler, T., Harper, R., & Wright, W. (2008). Stories in geotime. Information Visualization, 7, 3–17.

Fischer, G. (2001). User modeling in human–computer interaction. User Modeling and User-Adapted Interaction, 11, 65–86.

Fuxman, A., Tsaparas, P., Achan, K., & Agrawal, R. (2008). *Using the wisdom of the crowds for keyword generation*. In Proceedings of the 17th International Conference on World Wide Web. New York: ACM.

Gallego, M.D., Luna, P., & Bueno, S. (2008). User acceptance model of open source software. Computers in Human Behavior, 24, 2199–2216.

Gandy, L., Rahimi, S., & Gupta, B. (2005). *A modified competitive agglomeration for relational data algorithm*. In Fuzzy Information Processing Society, 2005. NAFIPS 2005. Annual Meeting of the North American. Washington, DC: IEEE Computer Society.

Girvan, M., & Newman, M.E. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences of United States of America, 99, 7821–7826.

Godoy, D., & Amandi, A. (2005). User profiling in personal information agents: A survey. The Knowledge Engineering Review, 20, 329–361.

Görg, C., Liu, Z., & Stasko, J. (2013). Reflections on the evolution of the Jigsaw visual analytics system. Information Visualization, 0:(0), 1–11.

Gotz, D., & Zhou, M.X. (2009). Characterizing users' visual analytic activity for insight provenance. Information Visualization, 8, 42–55.

Gove, R. P. (2011). Understanding scientific literature networks: Case study evaluations of integrating visualizations and statistics (Doctoral dissertation).

Greenhalgh, T., Robert, G., Macfarlane, F., Bate, P., & Kyriakidou, O. (2004). Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Quarterly, 82, 581–629.

Guo, C., Liu, Y., Shen, W., Wang, H.J., Yu, Q., & Zhang, Y. (2009). *Mining the web and the internet for accurate ip address geolocations*. In INFOCOM 2009, IEEE. Washington, DC: IEEE Computer Society.

Harris, M., & Butterworth, G. (2012). Developmental psychology: A student's handbook. Psychology Press.

He, D., Göker, A., & Harper, D.J. (2002). Combining evidence for automatic web session identification. Information Processing & Management, 38, 727–742.

Heer, J., Card, S.K., & Landay, J.A. (2005). *Prefuse: A toolkit for interactive information visualization*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM.

Hindus, D., Ackerman, M.S., Mainwaring, S., & Starr, B. (1996). *Thunderwire: A field study of an audio-only media space*. In Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work. New York: ACM.

Hu, B., Zhang, Y., Chen, W., Wang, G., & Yang, Q. (2011). *Characterizing search intent diversity into click models*. In Proceedings of the 20th International Conference on World Wide Web. New York: ACM.

Huang, Z., Ng, J., Cheung, D.W., Ng, M.K., & Ching, W.K. (2001). *A cube model for web access sessions and cluster analysis*. In Proceedings of WEBKDD. Heidelberg:Springer-Verlag Berlin.

Kamvar, M., Kellar, M., Patel, R., & Xu, Y. (2009). *Computers and iphones and mobile phones, oh my!: A logs-based comparison of search users on different devices*. In Proceedings of the 18th International Conference on World Wide Web. New York: ACM.

Kang, Y.A., Gorg, C., & Stasko, J. (2011). How can visual analytics assist investigative analysis? Design implications from an evaluation. IEEE Transactions on Visualization and Computer Graphics, 17, 570–583.

Karrer, B., & Newman, M.E. (2011). Stochastic blockmodels and community structure in networks. Physical review E, 83, 016107.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 79–86.

Kumar, R., & Tomkins, A. (2010). *A characterization of online browsing behavior*. In Proceedings of the 19th International Conference on World Wide Web. New York: ACM.

Lee, U., Liu, Z., & Cho, J. (2005). *Automatic identification of user goals in web search*. In Proceedings of the 14th International Conference on World Wide Web. New York: ACM.

Lin, J., & Wilbur, W.J. (2009). Modeling actions of PubMed users with n-gram language models. Information Retrieval, 12, 4), 487–503.

Linton, F., Joy, D., & Schaefer, H.P. (1999). Building user and expert models by long-term observation of application usage. Proceedings of the Seventh International Conference on User modeling, 129–138, Springer-Verlag New York.

Liu, N., Liu, Y., & Wang, X. (2010). *Data logging plus e-diary: Towards an online evaluation approach of mobile service field trial*. In Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services. New York: ACM.

Lu, Z., Wilbur, W.J., McEntyre, J.R., Iskhakov, A., & Szilagyi, L. (2009). *Finding query suggestions for PubMed*. In American Medical Informatics Association.

Mackinlay, J., Hanrahan, P., & Stolte, C. (2007). Show me: Automatic presentation for visual analysis. IEEE Transactions on Visualization and Computer Graphics, 13, 1137–1144.

Manning, C.D., & Schütze, H. (1999). Foundations of statistical natural language processing. (Vol. 999). Cambridge: MIT press.

Mobasher, B., Cooley, R., & Srivastava, J. (1999). *Creating adaptive web sites through usage-based clustering of URLs*. In Proceedings, 1999 Workshop on Knowledge and Data Engineering Exchange(KDEX'99). Washington, DC: IEEE Computer Society.

Nasraoui, O., Frigui, H., Joshi, A., & Krishnapuram, R. (1999). *Mining web access logs using relational competitive fuzzy clustering*. In Proceedings of the Eight International Fuzzy Systems Association World Congress. Washington, DC: IEEE Computer Society.

Newman, M.E. (2004). Fast algorithm for detecting community structure in networks. Physical review E, 69(6), 066133.

Norris, J.R. (1998). Markov chains (Vol. 2). New York: Cambridge University Press.

Osiński, S., Stefanowski, J., & Weiss, D. (2004). Lingo: Search results clustering algorithm based on singular value decomposition Intelligent information processing and web mining (pp. 359–368). Berlin, Heidelberg: Springer.

Poese, I., Uhlig, S., Kaafar, M.A., Donnet, B., & Gueye, B. (2011). IP geolocation databases: Unreliable? ACM SIGCOMM Computer Communication Review, 41, 53–56.

Rieman, J. (1993). *The diary study: A workplace-oriented research tool to guide laboratory efforts*. In Proceedings of the INTERACT'93 and CHI'93 Conference on Human Factors in Computing Systems. New York: ACM.

Maimon, O., & Rokach, L. (Eds.). (2005). Clustering methods. Data mining and knowledge discovery handbook (Vol. 2, pp. 321–352). New York: Springer.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Schmid, B. (2012). An investigation of activity logging methods in user studies. Diploma thesis.

Shahabi, C., & Banaei-Kashani, F. (2003). Efficient and anonymous web-usage mining for web personalization. INFORMS Journal on Computing, 15, 123–147.

Shahabi, C., Zarkesh, A.M., Adibi, J., & Shah, V. (1997). *Knowledge discovery from users web-page navigation*. In Proceedings. Seventh International Workshop on Research Issues in Data Engineering, 1997. Washington, DC: IEEE Computer Society.

Silva, C.T., Anderson, E., Santos, E., & Freire, J. (2011). *Using vistrails and provenance for teaching scientific visualization*. Computer Graphics Forum (Vol. 30, No. 1, pp. 75–84).

Song, Q., & Shepperd, M. (2006). Mining web browsing patterns for E-commerce. Computers in Industry, 57, 622–630.

Song, Y., Ma, H., Wang, H., & Wang, K. (2013). *Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance*. In Proceedings of the 22nd International Conference on World Wide Web. New York: ACM.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsletter, 1, 12–23.

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: supporting investigative analysis through interactive visualization. Information Visualization, 7, 118–132.

Stephens, M. (1982). A question of generalizability. Theory & Research in Social Education, 9, 75–89.

Stolte, C., Tang, D., & Hanrahan, P. (2002). Polaris: A system for query, analysis, and visualization of multidimensional relational databases. IEEE Transactions on Visualization and Computer Graphics, 8, 52–65.

Suryavanshi, B.S., Shiri, N., & Mudur, S.P. (2005). *An efficient technique for mining usage profiles using relational fuzzy subtractive clustering*. In Proceedings. International Workshop on Challenges in Web Information Retrieval and Integration, 2005. Washington, DC: IEEE Computer Society.

Synnestvedt, M.B., & Chen, C. (2005). *Design and evaluation of the tightly coupled perceptual-cognitive tasks in knowledge domain visualization*. In Proceedings of the 11th International Conference on Human-Computer Interaction (HCI International 2005). New York: ACM.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523–538.

Viegas, F.B., Wattenberg, M., Van Ham, F., Kriss, J., & McKeon, M. (2007). Manyeyes: A site for visualization at internet scale. IEEE Transactions on Visualization and Computer Graphics, 13, 1121–1128.

Wan, M., Jönsson, A., Wang, C., Li, L., & Yang, Y. (2012). Web user clustering and web prefetching using random indexing with weight functions. Knowledge and Information Systems, 33, 89–115.

Ward, Jr., J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58, 236–244.

Xie, Y., & Phoha, V.V. (2001). *Web user clustering from access log using belief function*. In Proceedings of the 1st International Conference on Knowledge Capture. New York: ACM.

Xu, J., & Liu, H. (2010). *Web user clustering analysis based on KMeans algorithm*. In 2010 International Conference on Information Networking and Automation (ICINA). Washington, DC: IEEE Computer Society.

Yan, T.W., Jacobsen, M., Garcia-Molina, H., & Dayal, U. (1996). From user access patterns to dynamic hypertext linking. Computer Networks and ISDN Systems, 28, 1007–1014.

Yuan, X., Chen, C., Zhang, X., Avery, J., & Xu, T. (2013). Effects of domain knowledge on user performance and perception in a knowledge domain visualization system *Design, User Experience, and Usability. Web, Mobile, and Product Design* (pp. 601–610). Springer Berlin Heidelberg.

Zhang, Z., & Nasraoui, O. (2006). *Mining search engine query logs for query recommendation*. In Proceedings of the 15th International Conference on World Wide Web. New York: ACM.

Zhao, Q., Hoi, S.C., Liu, T.Y., Bhowmick, S.S., Lyu, M.R., & Ma, W.Y. (2006). *Time-dependent semantic similarity measure of queries using historical click-through data*. In Proceedings of the 15th International Conference on World Wide Web. New York: ACM.

Appendix: Clustering on Sampled Sessions

In this Appendix, we sampled 10% of all 251,954 sessions, and did a clustering on the sample. We found that in the sampled sessions, there is not much difference between different clusters of behavior patterns. In other words, the clusters discovered reveal sub-pattern of a major behavior pattern, rather than higher-level distinct patterns. This is because some behavior pattern has generated far more number of sessions than the others. In this case, taking every session equally in clustering might result a biased result towards the major behavior pattern.

More specifically, first, we took an overview of the entire data set by comparing the number of sessions generated in different clusters of session aggregations (for the entire data set). Second, we randomly sampled 10% of all sessions and performed clustering in the same way as we did in our main analysis.

TABLE 1. Number of sessions of each cluster (based on IP addresses)

Cluster	Total number of sessions
#1: entry-level behavior pattern	10,960
#2: passable-level behavior pattern	15,968
#3: accomplished-level behavior pattern	225,020

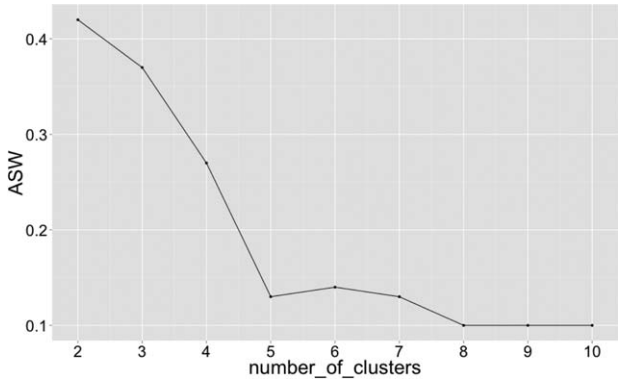

FIG. 1. Average Silhouette Width of clusters on 10% sampled sessions.

1. Overview of entire data set in terms of session distribution

We took an overview of the entire data set by comparing the number of sessions generated by the three clusters of session aggregations (for the entire data set) in TABLE 1. We could observe that cluster #3 generated far more number of sessions than cluster #1 and cluster #2. This suggests that the number of sessions are distributed extremely uneven across different clusters. Therefore, if we go backwards and cluster on sessions, the result would skew towards the one major behavior pattern, neglecting the rest patterns. This is supported in the next part of the analysis.

2. Clustering on sampled sessions

In order to have a better idea of what the clusters would be if we cluster on all sessions instead of session aggregations, we randomly sampled 10% of all 251,954 sessions and clustered on the sampled sessions.

2.1. Average Silhouette Width (ASW) of the Sampled Data. The ASW of the clustering at different number of clusters is depicted in FIG. 1. The clustering result of the 10% sampled sessions is not as good as our previous results of clustering on session aggregations. The average silhouette width is around 3.7 to 4.2, compared to previous result 4.2 to 5.8 (k=2,3).

2.2. Hierarchical Clustering of the Sampled Data. The MDS map of the hierarchical clustering is depicted in FIG. 2. The MDS map shows that the clustering doesn't

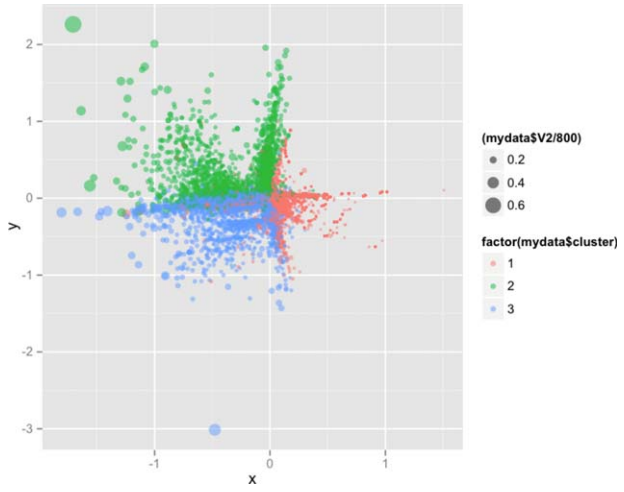

FIG. 2. MDS of clustering on 10% sampled sessions. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

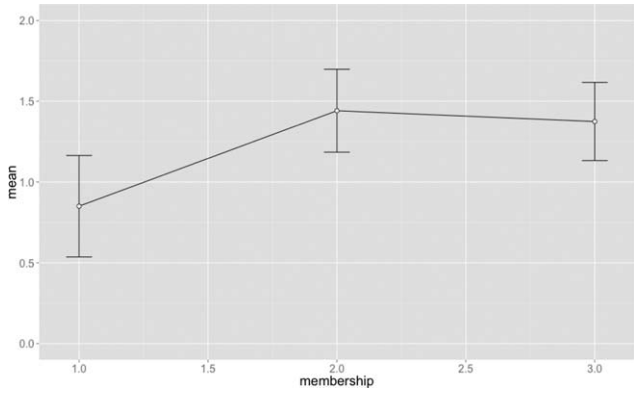

FIG. 3. Bar-plot of logarithm value of number of events of each cluster.

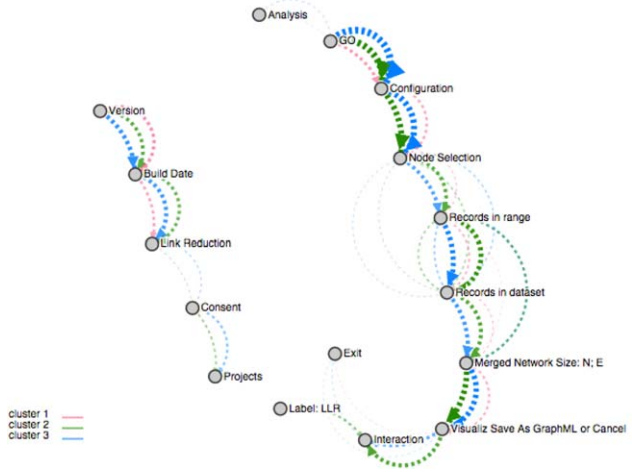

FIG. 4. Behavior patterns of 3 clusters of sampled 10% sessions. [Color figure can be viewed at wileyonlinelibrary.com]

seem to be able to separate the cluster very well, as shown in the MDS map in FIG.2. The red cluster is overlapped under green and blue clusters.

2.3. Behavior Patterns of the Sampled Data. When taking a closer look at features of each cluster, we find that the differences between clusters are narrowed, compared to previous results.

First, the numbers of events within each cluster are not significantly different, compared to previous results. This can be observed from FIG. 3.

Second, when we compare the behavior patterns, we can see that all three clusters, that is cluster #1 (entry-level: red), #2 (accomplished-level: green), and #3 (passable-level: blue) reached the event of "interaction," but with different probabilities (more likely for cluster #2 and cluster #3, but less likely for cluster #1) as in FIG. 4. But in our major analysis, the behavior patterns are very different (please refer to FIG. 9. in our submission). This means that, when we use sessions instead of session aggregations for clustering, the difference between clusters is not as vast as that between session aggregations, although there are still differences. This may be due to the fact that when we treat all sessions equally and sample on them randomly, it is possible that we sampled more "accomplished-level sessions" than "entry-level sessions," since originally there are more experts than beginners in our data, and experts tend to have more sessions and more number of events. So in our sample, we are actually trying to divide the advanced-level sessions into further clusters. This may explain why the behavior patterns are not so different from each other.

If we extrapolate the above findings into the whole data set of 251,954 sessions, the result should be similar. The sessions from advanced patterns would take more share than entry-level patterns, which leads to clusters similar to each other. But when we cluster on sessions aggregations, it is similar to a stratified sampling, which means that we take similar sessions (from the same IP) as one unit for clustering.

DOI: 10.1002/asi