Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

Navigation Space Based Intranet Usability Analysis

Peter GÉCZY, Noriaki IZUMI, Shotaro AKAHO, Kôiti HASIDA
National Institute of Advanced Industrial Science and Technology
Tsukuba and Tokyo, Japan

Abstract— **Usability is a vital quality factor of electronic environments. Increasing complexity of web based platforms and mining large data volumes constantly rise demands on effective usability analysis tools. The article presents a novel formalism allowing examination of usability characteristics with the spectrum of metrics. The approach was applied to the usability analysis of a large corporate Intranet. Majority of the users were knowledge workers. Important usability features have been revealed. The knowledge workers efficiently utilized only a minor portion of available Intranet resources and exhibited strong pattern formation tendency. The frequently repeating browsing patterns were generally easily executable.**

I. INTRODUCTION

"Nobody has really looked at productivity in white collar work in a scientific way." (Peter Drucker) [1]. Absence of the scientific evidence regarding the knowledge worker productivity, efficiency, and their relevant metrics have been at the center of recent managerial discourse [2]. Exploration of human dynamics [3],[4] and usability of electronic environments have been attracting significant attention in a corporate sector; ranging from mobile [5] to web based e-commerce [6].

Observer and protocol based usability studies provide valuable behavioral data [7], however, they are human resource demanding and time consuming. Partial automation of the behavioral studies has been achieved using attention tracking systems (eye tracking devices) [8] and self-administration of adaptively modified tests [9]. Combinations of the human observer, third-party ratings, and software agents have also been proposed [10]. Substantial activity in the usability research has led to a wide range of models and metrics. Only limited unifying attempts have been made [11]. Recent trends in the usability analysis target the 'non-invasive' automated approaches that permit on-the-fly massive data mining. Statistical and cluster analysis methods have been applied in this area [12]. However, the large data volumes convey extra challenges. On-the-fly processing and adaptive functionality of the systems require a considerable computing power. It is desirable to keep the computational complexity and the resource requirements of the analysis minimal.

We propose a novel approach to the usability analysis utilizing a navigation space construct that effectively captures the dimensions of human interactions, and is efficiently implementable. The navigation space encompasses both topological and temporal characteristics of a human behavior, allows categorization of the users with similar browsing behavior, and permits the projections of the user clusters to various interest domains. The construct was applied to the usability analysis of a large corporate Intranet.

II. APPROACH FORMULATION

We introduce the basic approach to the Intranet usability analysis utilizing the navigation space construct. Definitions are accompanied by the intuitive explanations that help us better understand the concept at a more formal level.

The click-stream sequences [13] of page transitions are divided into sessions, and sessions are further divided into subsequences. The division is done with respect to the user activity and inactivity. Consider the conventional time-stamp click-stream sequence of the following form: $\{(p_i, t_i)\}_i$, where p_i denotes the visited page URL_i at the time t_i. For the purpose of analysis this sequence is converted into the form: $\{(p_i, d_i)\}_i$ where d_i denotes the delay between the consecutive views $p_i \rightarrow p_{i+1}$. The user browsing activity $\{(p_i, d_i)\}_i$ is divided into subelements according to the periods of inactivity d_i satisfying certain criteria.

Definition 1: (Browsing Session, Subsequence, Train) *Let $\{(p_i, d_i)\}_i$ be a sequence of pages p_i with delays d_i between consecutive transitions $p_i \rightarrow p_{i+1}$.*
Browsing session *is a sequence $B = \{(p_i, d_i)\}_i$ where each $d_i \leq T_B$. Length of the browsing session is $|B|$. Browsing session is often referred to simply as a* **session**.
Subsequence *of an individual browsing session B is a sequence $S = \{(p_i, dp_i)\}_i$ where each delay $dp_i \leq T_S$, and $\{(p_i, dp_i)\}_i \subset B$. The length of subsequence is $|S|$.*
A browsing session B thus consists of a **train** *of subsequences S_i separated by inactivity delays ds_i, $B = \{(S_i, ds_i)\}_i$.*

Sessions delineate tasks of various complexities users undertake on the Intranet. Subsequences correspond to the session subgoals; e.g. subsequence S_1 is a login, S_2 – a document download, S_3 – a search for internal resource, etc.

Important issue is determining the appropriate values of T_B and T_S that segment the user activity into sessions and subsequences. The former research [14] indicated that the student browsing sessions last on average 25.5 minutes. However, we adopt the average maximum attention span of 1 hour as a value for T_B. If the user's browsing activity was followed by a period of inactivity greater than 1 hour, it is considered a single session, and the following activity comprises the next session.

Value of T_S is determined dynamically and computed as an average delay in the browsing session: $T_S = \frac{1}{N}\sum_{i=1}^{N} d_i$. If the delays between page views are short, it is useful to bound the value of T_S from below. This is preferable in environments with frame-based and/or script generated pages where numerous logs are recorded in a rapid transition. Since our situation contained both cases, we adjusted the value of T_S by bounding it from below by 30 seconds:

$$T_S = max\left(30, \frac{1}{N}\sum_{i=1}^{N} d_i\right). \quad (1)$$

Using these primitives we define the navigation space and subspace as follows.

Definition 2: (Navigation Space and Subspace)
Navigation space *is a triplet* $\mathcal{G} = (\mathcal{P}, \mathcal{B}, \mathcal{S})$ *where* $\mathcal{P}$ *is a set of points (e.g. URLs),* $\mathcal{B}$ *is a set of browsing sessions, and* $\mathcal{S}$ *is a set of subsequences.*
Navigation subspace *of* $\mathcal{G}$ *is a space A = (D,H,K) where* $D \subseteq \mathcal{P}$, $H \subseteq \mathcal{B}$, *and* $K \subseteq \mathcal{S}$; *denoted as* $A \subseteq \mathcal{G}$.

The navigation space can be divided into subspaces based on the nature of detected or defined sequences. For example, a human navigation space consists of human generated sequences, and a machine navigation space may contain only the machine generated sequences. Different spaces may have distinctly different characteristics.

Another important aspect is to observe where the user actions are initiated and terminated. That is, to identify the starting and ending points of subsequences, as well as single user actions.

Definition 3: (Starter, Attractor, Singleton)
Let $\mathcal{G} = (\mathcal{P}, \mathcal{B}, \mathcal{S})$ *be a navigation space and* $B = \{(S_i, ds_i)\}_i^M$, $B \in \mathcal{B}$, *be a browsing session, and* $S = \{(p_k, dp_k)\}_k^N$, $S \in \mathcal{S}$, *be a subsequence.*
Starter *is the first point of an element of subsequence or session with length greater that 1, that is,* $p_1 \in \mathcal{P}$ *such that there exist* $B \in \mathcal{B}$ *or* $S \in \mathcal{S}$ *where* $|B| > 1$ *or* $|S| > 1$ *and* $(p_1, d_1) \in B$ *or* $(p_1, dp_1) \in S$.
Attractor *is the last point of an element of subsequence or session with length greater that 1, that is,* $p_N \in \mathcal{P}$ *or* $p_M \in \mathcal{P}$ *such that there exist* $B \in \mathcal{B}$ *or* $S \in \mathcal{S}$ *where* $|B| > 1$ *or* $|S| > 1$ *and* $(p_M, d_M) \in B$ *or* $(p_N, dp_N) \in S$.
Singleton *is a point* $p \in \mathcal{P}$ *such that there exist* $B \in \mathcal{B}$ *or* $S \in \mathcal{S}$ *where* $|B| = 1$ *or* $|S| = 1$ *and* $(p, d) \in B$ *or* $(p, dp) \in S$.

The starters, attractors, and singletons are important points in a navigation space. The starters refer to the starting navigation points of user actions, whereas the attractors denote the users' targets. The singletons relate to the single user actions such as use of hotlists (e.g. history or bookmarks) [15]. Note that a single point p can be the starter, attractor, and/or singleton.

We can formulate behavioral abstractions simply as pairs of starters and attractors. Then it is equally important to observe the connecting elements of transitions from one task (or subtask) to the other.

Definition 4: (SE Elements, Connectors)
Let $B = \{(S_i, ds_i)\}_i$ *be a browsing session and* $S_i = \{(p_{ik}, dp_{ik})\}_k^N$ *and* $S_{i+1} = \{(p_{i+1l}, dp_{i+1l})\}_l^M$ *be consecutive subsequences* $S_i \to S_{i+1}$ *of* B.
SE element *(start-end element) of a subsequence* S_i *is a pair* $SE_i = (p_{i1}, p_{iN}) = (\psi_s, \psi_e)$.
Connector *of subsequences* S_i *and* S_{i+1} *is a pair* $C_i = (p_{iN}, p_{i+1,1}) = (\eta_s, \eta_e)$.

The SE elements outline the higher order abstractions of user subgoals. Knowing the starting point, users can follow various navigational pathways to the target. Focusing on the starting and ending points of user actions eliminates the variance of the navigational choices. The connectors indicate the links between elemental browsing patterns. This enables us to observe formation of more complex behavioral patterns as interconnected sequences of elemental patterns.

Definition 5: (Range)
Let $\mathcal{G} = (\mathcal{P}, \mathcal{B}, \mathcal{S})$ *be a navigation space.* **Range** $r_p = < r_{min}^{(p)}, r_{max}^{(p)} >$ *of a point* p *is the minimum and maximum length of remaining sequences from the given position of* $(p, d) \in B$ *or* $(p, dp) \in S$ *to either starter or attractor points. Range of a singleton is* **0**.

The range is an important point characteristic. It can be intuitively perceived as an extent of required navigational transitions in order to reach the nearest attractor or starter from a given point in a navigation space.

Combining the defined primitives we can express the usability of specific elements in the form of a vector.

Definition 6: (Usability Vector)
Let $\mathcal{G} = (\mathcal{P}, \mathcal{B}, \mathcal{S})$ *be a navigation space.* **Usability vector** μ *is the normalized vector* $\mu = ||(\mu_c, \mu_{r_{min}}, \mu_{r_{max}})||_2$ *, with respect to* l_2 *norm. Vector coordinates for points, SE elements, and connectors are defined as follows:*

Point $p \in \mathcal{P}$**:**

$\mu_c = \frac{c}{max_p(c)}$, *where* c *is the number of occurrences of* p,
$\mu_{r_{min}} = \frac{r_{min}^{(p)}}{max_p(r_{min}^{(p)})}$, $r_{min}^{(p)}$ *is the minimum range of* p,
$\mu_{r_{max}} = \frac{r_{max}^{(p)}}{max_p(r_{max}^{(p)})}$, $r_{max}^{(p)}$ *is the maximum range of* p.

SE Element SE_i**:**

$\mu_c = \frac{c}{max_{SE_i}(c)}$, *where* c *is number of occurrences of* SE_i,
$\mu_{r_{min}} = \frac{SE_{i,r_{min}}}{max_{SE_i}(SE_{i,r_{min}})}$, *and* $SE_{i,r_{min}} = \frac{r_{min}^{(\psi_s)} + r_{min}^{(\psi_e)}}{2}$,
$\mu_{r_{max}} = \frac{SE_{i,r_{max}}}{max_{SE_i}(SE_{i,r_{max}})}$, *and* $SE_{i,r_{max}} = \frac{r_{max}^{(\psi_s)} + r_{max}^{(\psi_e)}}{2}$.

Connector C_i**:**

$\mu_c = \frac{c}{max_{C_i}(c)}$, *where* c *is number of occurrences of* C_i,
$\mu_{r_{min}} = \frac{C_{i,r_{min}}}{max_{C_i}(C_{i,r_{min}})}$, *and* $C_{i,r_{min}} = \frac{r_{min}^{(\eta_s)} + r_{min}^{(\eta_e)}}{2}$,
$\mu_{r_{max}} = \frac{C_{i,r_{max}}}{max_{C_i}(C_{i,r_{max}})}$, *and* $C_{i,r_{max}} = \frac{r_{max}^{(\eta_s)} + r_{max}^{(\eta_e)}}{2}$.

The above formulation enables usability evaluation of not only the navigation points, but also the users' behavioral abstractions and their connecting actions. This allows more complete usability analysis.

Since the number of occurrences, and the minimum and maximum ranges may vary substantially in values, they are scaled to 1, in order to obtain relatively comparable measurements. The scaling is indicated by dividing their actual values by the maximum observed value for each particular component. The difference between points $p \in \mathcal{P}$, and SE elements and connectors (pairs of points), is simply averaging the minimum and maximum range values of the pair.

The normalization of a usability vector μ can be carried out with respect to various norms, however, we consider l_2 norm. It leads to the directional cosines of the scaled numbers of occurrences, and the minimum and maximum ranges, and permits relevant numerical evaluation, as well as effective usability visualization.

Rigorous approach demands also elucidation of temporal characteristics of navigation spaces related to durations and delays, however, this study focuses primarily on exploring topological features.

III. Intranet and Data

Data used in this work was a one year period Intranet web log data of The National Institute of Advanced Industrial Science and Technology (Table I). The majority of users are skilled knowledge workers. Intranet web portal had a load balancing architecture comprising of 6 servers providing extensive range of web services and documents vital to the organization. The Intranet services support managerial, administration and accounting processes, research cooperation with industry and other institutes, databases of research achievements, resource localization and search, attendance verification, and also numerous bulletin boards and document downloads. The institution has a number of branches at various locations throughout the country, thus certain services are decentralized. The size of a visible web space was approximately 1 GB. The invisible web size was considerably larger, but difficult to estimate due to the distributed architecture and constantly changing back-end data.

TABLE I

Primary information about data used in the study.

Data Volume	∼60 GB
Average Daily Volume	∼54 MB
Number of Servers	6
Number of Log Files	6814
Average File Size	∼9 MB
Time Period	3/2005 - 4/2006

Daily traffic was substantial and so was the data volume. Information summary of the data is presented in Table I. It is important to note that the data was incomplete. Although some days were completely represented, every month there were missing logs from specific servers. The server side logs also suffered a data loss due to the caching and proxing. However, because of the large data volume, the missing data only marginally affected the analysis. Web servers run the open source Apache server software and the web log data was in the combined log format without referrer.

IV. Navigation Space Extraction

Starting with the setup description we present the data preprocessing and cleaning. Then we proceed to the session and subsequence extraction. Machine generated traffic was still present even after initial data cleaning. Extraction of a human navigation subspace and filtering of non-human traffic is described during the subsequence extraction.

Setup. Extraction and analysis of a knowledge worker navigation space from Intranet web logs was performed on Linux setup with MySQL database as a data storage engine for preprocessed and processed data. Analytic and processing routines were implemented in various programming languages and optimized for high performance.

Preprocessing. Data fusion of the web logs from 6 servers of a load balanced Intranet architecture was performed at the preprocessing level. Data was largely contaminated by the logs from automatic monitoring software and required filtering. During the initial filtering phase the logs from software monitors, invalid requests, web graphics, style sheets, and client-side scripts were eliminated. The access logs from scripts, downloadable and syndicated resources, and documents in various formats were preserved. Information was structured according to the originating IP address, complete URL, base URL, script parameters, date-time stamp, source identification, and basic statistics. Clean raw data was logged into a database and appropriately linked.

TABLE II

Processed web log data statistics.

Log Records	315 005 952
Clean Log Records	126 483 295
Unique IP Addresses	22 077
Unique URLs	3 015 848
Scripts	2 855 549
HTML Documents	35 532
PDF Documents	33 305
DOC Documents	4 385
Others	87 077

Approximately 40.15% of the original log records remined after the initial filtering (see Table II). Major access to Intranet resources was via scripts (94.68%). Only relatively minor portions of accessible resources were static HTML documents (1.18%), PDF documents (1.1%), DOC documents (0.15%), and others (2.89%), such as downloadable software, updates, spreadsheets, syndicated resources, etc. Detected IP address space (22077 unique IPs) consisted of both statically and dynamically assigned IP addresses.

Session Extraction. Preprocessed and databased Apache web logs (in combined log format) did not contain referrer information. The click-stream sequences were reconstructed by ordering the logs originating from the unique IP addresses according to the time-stamp information. Ordered log sequences from the specific IP addresses were divided into the browsing sessions as described in Definition 1. Session divisor was the predetermined user inactivity period ds_i greater than $T_B = 1\ hour$.

It is noticeable (see Table III) that the user sessions on the corporate Intranet were on average longer (appx. 48.5 minutes) than those of the students (appx. 25.5 minutes) reported in [14]. The average number of 156 sessions per IP address, and a large variation in the maximum and minimum number of sequences from distinct IP addresses, indicate that association

TABLE III

OBSERVED BASIC SESSION DATA STATISTICS.

Sessions	3 454 243
Unique Sessions	2 704 067
Average Sessions per Day	9 464
Average Session Length	36 [transitions]
Average Session Duration	2 912.23 [s] (48 min 32 sec)
Average dp_i Delay per Session	81.55 [s] (1 min 22 sec)
Average Sessions per IP Address	156
Maximum	1 553
Minimum	1

of users with distinct IP addresses is relevant only for the registered static IP addresses. Large number (3492) of single sessions only originated from distinct IP addresses due to DHCP use.

Extraction of Subsequences and Human Navigation Space. Each detected session was analyzed for the subsequences as defined in Definition 1. Lower bound of 30 seconds for the separating inactivity period dp_i was proper.

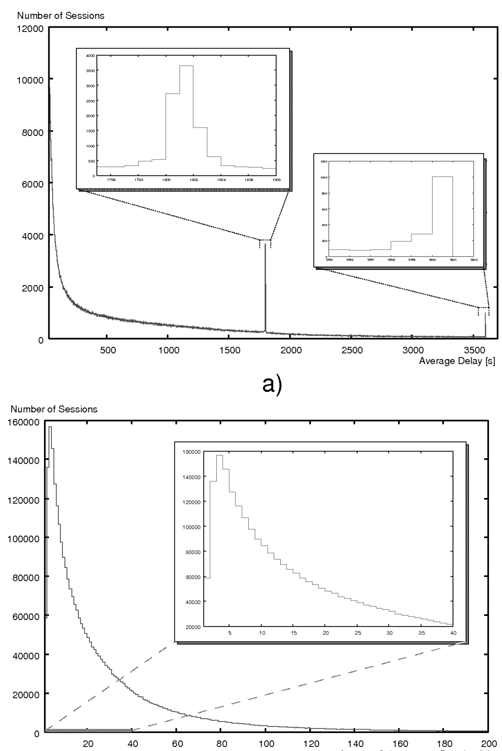

Fig. 1. Histograms: **a)** average delay between subsequences in sessions, **b)** average subsequence duration. There are noticeable spikes in chart **a)** around 1800 seconds (30 minutes) and 3600 seconds (1 hour). The detailed view is displayed in subcharts. Temporal variation of spikes corresponds to the peak average subsequence duration in chart **b)**. The spikes with relatively accurate delays between subsequences are due to machine generated traffic.

It has been observed that the sessions contained machine generated subsequences. As seen in the histogram of average delays between subsequences (Figure 1-a), there was a disproportionally large number of sessions with average delays between subsequences around 30 minutes and 1 hour. This is indicated by the spikes in Figure 1-a. The detailed view (subcharts of Figure 1-a) revealed that the variation in the average delay between subsequences was approximately ± 3 seconds. It correlates with the peak average subsequence duration (Figure 1-b). It is highly unlikely that human generated traffic would produce this precision.

The machine generated traffic contaminates the data and should be filtered, since we target primarily the human navigation space. We filtered two main groups of the machine generated subsequences: login subsequences and subsequences with delay periodicity around 30 minutes and 1 hour.

Every user is required to login into Intranet in order to access the services and resources. The login procedure involves validation and generates several log records with 0 delays. The records vary depending on whether the login was successful or unsuccessful. In both cases the log records and login related subsequences can be clearly identified and filtered.

The second group of machine generated traffic are the subsequences with periodicity of 30 minutes and 1 hour. Direct way of identifying these subsequences is to search for the sessions with only two subsequences having less than 1 second (or 0 second) duration (machines can generate requests fast and local Intranet servers are capable of responding within milliseconds) and the delay ds_i between subsequences within the intervals: 1800 and 3600 $\pm$ 3 seconds. It has been discovered that substantial number of such sessions contained relatively small number (170) of unique subsequences. Furthermore, these subsequences contained only 120 unique URLs. The identified subsequences and URLs were considered to be machine generated and filtered from the further analysis. Moreover, the subsequences with the SE elements containing identified URLs were also filtered.

TABLE IV

OBSERVED BASIC SUBSEQUENCE DATA STATISTICS.

Subsequences	7 335 577
Valid Subsequences	3 156 310
Filtered Subsequences	4 179 267
Unique Subsequences	3 547 170
Unique Valid Subsequences	1 644 848
Average Subsequences per Session	3
Average Subsequence Length	4.52 [transitions]
Average Subsequence Duration	30.68 [s]
Average ds_i Delay	388.46 [s] (6 min 28 sec)

Filtering of detected machine generated subsequences and their URLs significantly reduced the total number of subsequences - by 56.97% (from 7335577 to 3156310), as well as the number of unique subsequences - by 46.37% (from 3547170 to 1644848). Since the login sequences were also filtered, the number of subsequences per session decreased at least by 1. Reduction also occurred in the session lengths due to filtering of the identified invalid URLs. It is noticeable that the average subsequence duration (30.68 seconds) is approximately equal to the chosen lower bound for ds_i.

V. INTRANET USABILITY ANALYSIS

We present the Intranet usability analysis utilizing the navigation space construct. By analyzing the point characteristics together with behavioral abstractions and introduced usability metrics we infer several relevant observations. Analysis demonstrates usefulness of the navigation space formalism in elucidating usability features.

A. Starters, Attractors, and Singletons.

The point characteristics of a navigation space highlight the initial and terminal targets of knowledge worker activities, and also single-action behaviors.

TABLE V

STATISTICS FOR STARTERS, ATTRACTORS, AND SINGLETONS.

	Starters	Attractors	Singletons
Total	7 335 577	7 335 577	1 326 954
Valid	2 392 541	2 392 541	763 769
Filtered	4 943 936	4 943 936	563 185
Unique	187 452	1 540 093	58 036
Unique Valid	115 770	288 075	57 894

It is evident that the knowledge worker navigation space is substantially smaller, in this respect, than the observed complete navigation space (Table V). Reduction of the starters and attractors is approximately 67.4% ($7335577 \rightarrow 2392541$), and the singletons 57.56% ($1326954 \rightarrow 763769$).

Knowledge workers utilized relatively small spectrum of starting navigation points during their browsing. The set of starters, i.e. the initial navigation points of knowledge workers' (sub-)goals, was approximately 3.84% of the total navigation points. Initiation of actions from relatively small number of pages could be attributed to the several factors: a) small number of pages contained relevant links to transitional and/or target pages; b) knowledge workers were familiar with relatively small number of pages; c) knowledge workers found relatively small number of pages useful for initiating their actions; d) Intranet navigation structure contained relatively small number of pages with useful starting links.

Knowledge workers targeted relatively small number of available resources. Although the set of unique attractors, i.e. (sub-)goal targets, was approximately three times higher than the set of initial navigation points, it is still relatively minor portion of points in the navigation space (appx. 9.55% of the unique URLs). Knowledge workers aimed at relatively few resources, and generally displayed diminutive exploratory behavior. There might be various reasons for this outcome: a) knowledge workers had focused browsing interests; b) relatively small number of Intranet resources was found useful; c) Intranet navigation system concentrated on relatively small number of pages (hubs); d) navigation structure was infelicitous and discouraging for exploratory behavior.

Number of unique single user actions was minuscule. The single user actions, such as use of hotlists [15], followed by delays greater than 1 hour are represented by the singletons. The unique singletons accounted for only 1.92% of navigation points. If only a small number of starters and/or attractors was perceived useful, there is a possibility they were bookmarked and accessed directly in the following browsing experiences.

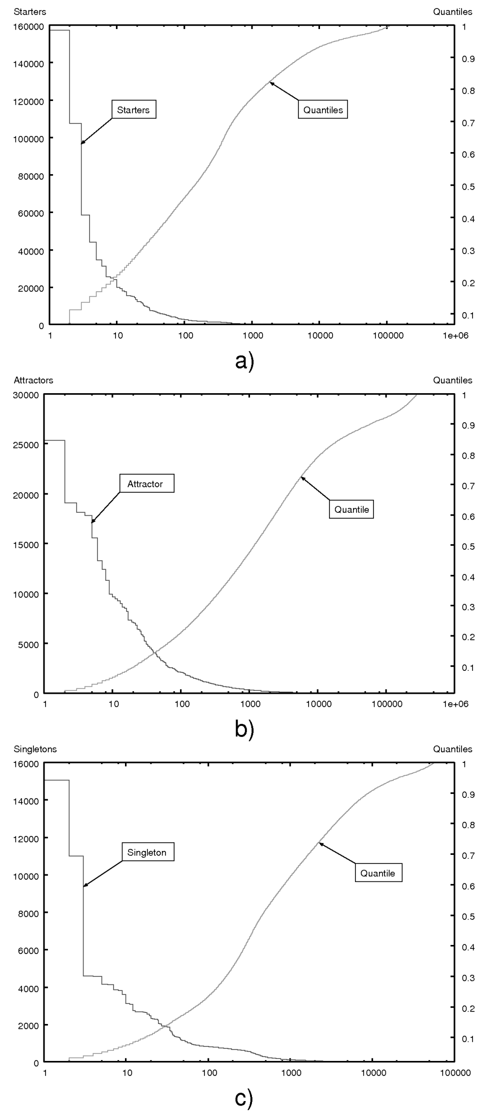

Fig. 2. Histograms and quantiles: **a)** starters, **b)** attractors, and **c)** singletons. Right y-axis contains a quantile scale. X-axis is in a logarithmic scale.

Few starters were frequently used. The histogram of starters (Figure 2-a left) indicates that higher frequency of occurrences is concentrated to relatively small number of elements. Approximately ten starters were very frequent and about one hundred were relatively frequent. The quantile analysis (Figure 2-a right) reveals that ten starters (appx. 0.0086% of the unique valid starters) accounted for about 20% of the total occurrences, and one hundred of them (appx. 0.086% of the unique valid starters) contributed to about 45%.

Few attractors were frequently used. As seen in Figure 2, small number of attractors was highly repetitive. The histogram curve (Figure 2-b left) shows that approximately fifty attractors were very frequent and about one thousand were relatively frequent. The quantile characteristics (Figure 2-b right) depict that fifty most frequent attractors (appx. 0.017% of the unique valid attractors) accounted for about 20% of the total observations, and one thousand attractors (appx. 0.35% of unique valid attractors) constituted about 48%.

Smaller number of starters (appx. one hundred) repeats substantially more frequently than the adequate number of attractors (appx. one thousand). It supports the hypothesis that *knowledge workers were generally more familiar with the starting browsing points rather than the targets.* In other words, they knew where to start and were familiar with the navigational path to the target.

Few singletons were frequently used. Analogous conditions are evident for singletons. The histogram chart (Figure 2-c left) depicts very frequent re-occurrence of approximately ten singletons and relatively frequent re-occurrence of approximately one hundred twenty singletons. The quantile projections (Figure 2-c right) highlight this fact. Ten most frequent singletons (appx. 0.017% of the unique valid singletons) accounted for about 20% of the total occurences, whereas one hundred twenty singletons (appx. 0.21% of the unique valid singletons) compounded to about 37% of the total occurences. Small number of very frequent single actions repeated with substantial delay suggests that only about ten navigation points were found substantially useful by the users to be included in their hotlists, or they required repetitive user's attention, such as attendance verification.

B. Ranges of Starters and Attractors.

The range of starters designates how near and/or far the knowledge workers go in their elemental browsing experiences. Analogously, the attractor range quantitatively describes how reachable the targets are, or how deeply in the Intranet structure they are embedded. Desirable situation wold be if the users were not required 'to go far' in order to reach their targets, or if they were assisted in faster reaching the targets.

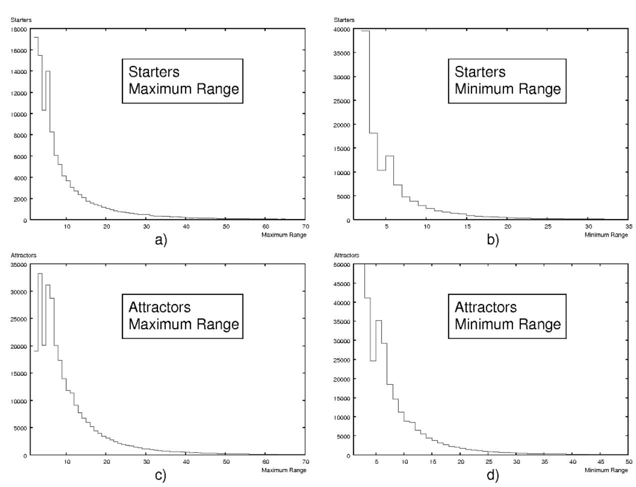

Fig. 3. Histograms: **a)** maximum ranges of starters, **b)** minimum range of starters, **c)** maximum range of attractors, and **d)** minimum range of attractors.

The observed frequent minimum ranges of starters were between 1 and 3 (Figure 3-b), and frequent maximum ranges of starters were in the interval <1,5> (Figure 3-a). Although the frequent attractor minimum ranges were also between 1 and 3 (with additional peaks at 5 and 6; see Figure 3-d), the more frequent maximum ranges of attractors were 3, 5, and 6 (Figure 3-c). This suggests that in order to reach certain targets, the knowledge workers were required to make more page transitions.

C. Usability Vector Analysis for Starters and Attractors.

Usability vectors encompass the combined range and frequency characteristics. Scaled and normalized values of the usability vector coordinates express how relatively close/far each starter or attractor is positioned to the observed number of occurrences, and the minimum and maximum ranges.

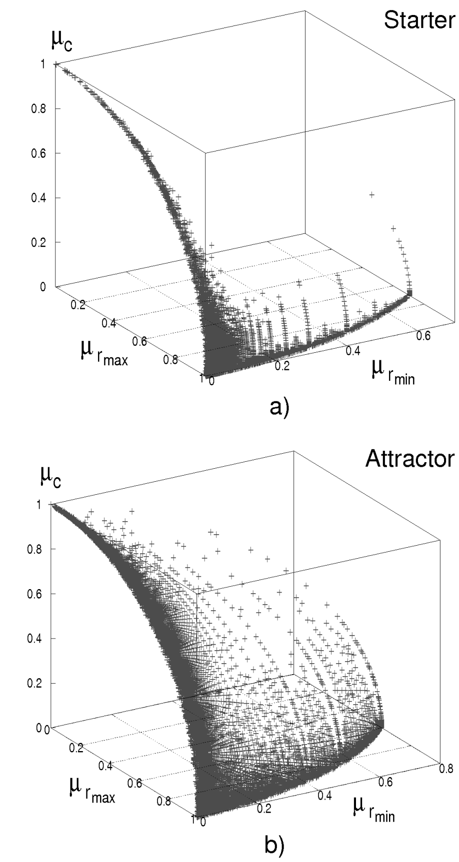

Fig. 4. Usability vector characteristics: **a)** starters, and **b)** attractors. All coordinates in 3D graphs depict normalized values.

Many infrequent starters and attractors were relatively difficult to reach. As displayed in Figure 4 there is a large number of starters and attractors with relatively small μ_c values, and relatively high values of $\mu_{r_{max}}$ and $\mu_{r_{min}}$. These are the infrequent elements (small μ_c) to which access was uneasy for most users. In order to reach these navigation points, users were required to make more transitions. It is a question whether this large amount of resources was infrequent because they were less important, or because of difficult accessibility. The former would be an acceptable condition, however, the later would call for potential improvement.

Frequent starters have relatively small minimum and maximum ranges. This is a desirable situation. Frequently accessed initial navigation points should be easily reachable. However, this condition holds only in the interval $\mu_c \in (0.9, 1)$. As μ_c values decreases below 0.8, the relative maximum range $\mu_{r_{max}}$ of starters rises beyond 0.6 (Figure 4-a). That is, relatively frequent starters were still more difficult to access.

Although the most frequent attractors have relatively small minimum and maximum ranges, there are many frequent attractors with higher minimum and maximum ranges. It is noticeable from Figure 4-b that a number of frequent attractors with $\mu_c > 0.8$ has the relative maximum range $\mu_{r_{max}} > 0.6$ and the minimum range $\mu_{r_{min}} > 0.2$. This situation should be improved.

D. SE Elements and Connectors.

These components serve as higher order abstractions of knowledge worker behavior. The SE elements represent the starting and ending points of subsequences, or the corresponding elemental patterns. The connectors delineate transitions between the pattern primitives, and thus formation of more complex patterns.

TABLE VI
STATISTICS FOR SE ELEMENTS AND CONNECTORS.

	SE Elements	Connectors
Total	7 335 577	3 952 429
Valid	2 392 541	2 346 438
Filtered	4 943 936	1 605 991
Unique	1 540 093	1 142 700
Unique Valid	1 072 340	898 896

There is a noticeable reduction of the SE elements and connectors in the knowledge worker navigation space (see Table VI). The number of SE elements decreased by 67.4% ($7335577 \rightarrow 2392541$) and connectors by 40.63% (from 3952429 to 2346438). Similarly, reduction is evident in the number of unique SE elements (30.37%: $1540093 \rightarrow 1072340$) and connectors (21.34%: $1142700 \rightarrow 898896$).

Small number SE elements and connectors was frequently repetitive. The histogram and quantile charts in Figure 5 depict re-occurrence of the SE elements and connectors. Approximately thirty SE elements and twenty connectors were very frequent (refer to the left histogram curves of Figure 5). These thirty SE elements (appx. 0.0028% of the unique valid SE elements) and twenty connectors (appx. 0.0022% of the unique valid connectors) accounted for about 20% of the total observations (see the right quantile curves of Figure 5).

Knowledge workers formed elemental and complex browsing patterns. Strong repetition of the SE elements indicates that knowledge workers often initiated their browsing actions from the same navigation point and targeted the same resource. This underlines the elemental pattern formation. Relatively

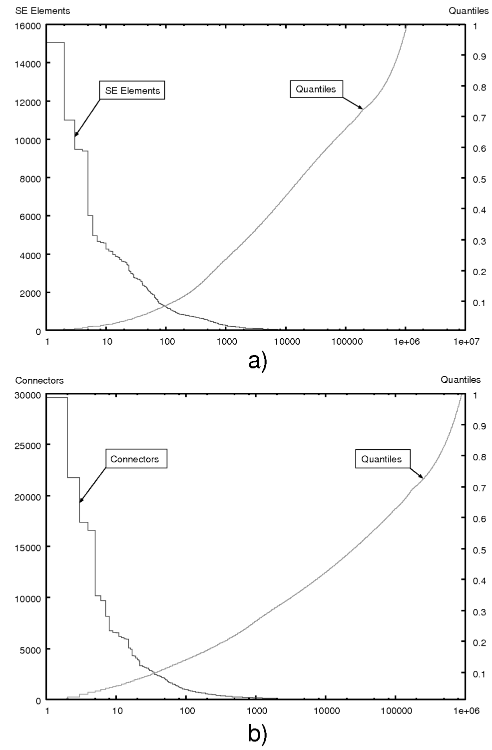

Fig. 5. Histograms and quantiles: **a)** SE elements, and **b)** connectors. Right y-axis contains a quantile scale. X-axis is in a logarithmic scale.

small number of elemental browsing patterns was frequently repeated. Re-occurrence of the connectors suggests that after completing a browsing sub-task, by reaching the desired target, they proceeded to the frequent starting point of the following sub-task(s). Frequently repeating elemental patterns interlinked with frequent transitions to other elemental sub-task highlights formation of more complex browsing patterns. Although the number of highly repetitive SE elements and connectors was small, knowledge workers exposed a spectrum of behavioral diversity in elemental as well as more complex behavioral patterns.

E. Usability Vector Analysis for SE Elements and Connectors.

Benefits of the presented formal approach extend to the usability metrics of knowledge worker behavioral abstractions and their connecting elements. The combined frequency and range attributes, scaled over averaged values of both elements in pair, delineate a relative ease or difficulty in executing elemental browsing patterns (represented by SE elements) and proceeding to other sub-tasks via connectors.

Frequent elemental and more complex browsing patterns required small number of transitions. This is a positive factor, since the often used browsing patterns and connecting transitions should be easily executable and reachable. It is noticeable from Figure 6 that a small number of frequent SE elements and connectors with a relatively high value of μ_c had small averaged relative minimum and maximum ranges $\mu_{r_{min}}$ and

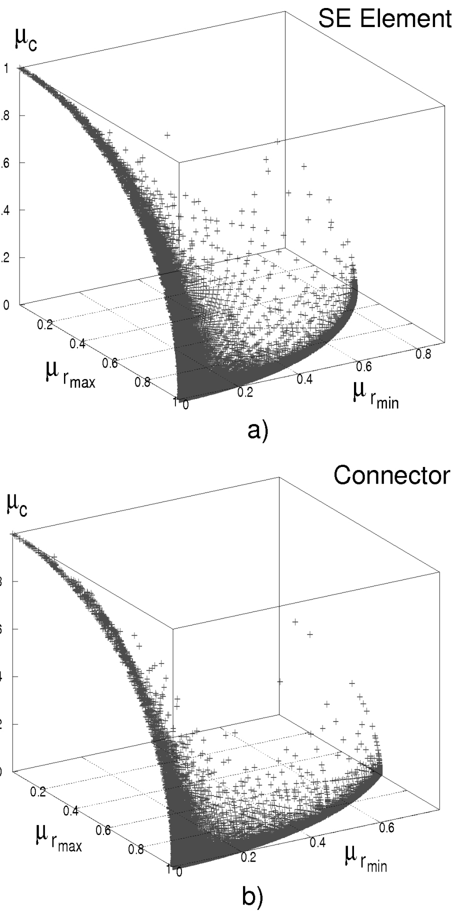

Fig. 6. Usability vector characteristics: **a)** SE elements, and **b)** connectors. All coordinates in 3D graphs depict normalized values.

$\mu_{r_{max}}$, respectively. However, this conditions holds for the SE elements and connectors with μ_c values greater than 0.9; then the relative maximum range $\mu_{r_{max}}$ rises.

Substantial number of infrequent SE elements and connectors was relatively difficult to execute and reach. The plots in Figure 6 show that there is a large number of infrequent SE elements and connectors—those with low values of μ_c. Major portion of them had relatively high maximum range $\mu_{r_{max}} > 0.8$, however, still relatively low minimum range $\mu_{r_{min}} < 0.2$. Smaller, but considerable, portion of the elements had the relative minimum range $\mu_{r_{min}} > 0.4$. It remains to be determined whether the execution complexity of these elemental browsing patterns was the major factor contributing to their infrequency, or they were simply occasional tasks.

Substantial number of frequent SE elements and connectors, e.g. those with the relative number of occurrences in the interval $\mu_c \in (0.6, 0.8)$, had high relative maximum range $\mu_{r_{max}} > 0.6$. This observation indicates that there is an ample room for *improvement of the Intranet portal whether in terms of the navigational structure, or utilization of advanced technologies, such as recommender systems,* that could shorten the maximum ranges and further assist knowledge workers.

VI. Conclusions and Future Work

Formal approach to usability analysis based on the navigation space construct has been introduced. Applicability of the framework may range from analyzing and modeling, to designing behaviorally centered web applications and services. The presented metrics enable effective evaluation of web components and behavioral characteristics of users. The usability analysis of a large corporate Intranet with substantial knowledge worker user base has been performed. Knowledge workers effectively utilized only small amount of available resources. The frequently accessed resources were relatively easy to locate. Knowledge workers had a significant tendency to form elemental and complex browsing patterns that were often reiterated. The frequent behavioral patterns were generally easily executable. Vast amount of resources has been occasionally accessed. The infrequently accessed resources required more complex navigation.

The future work targets clustering knowledge workers according to their usability and behavioral similarities. Results should be utilized for development of the next generation recommender systems and usability tools.

Acknowledgment

The authors would like to thank Tsukuba Advanced Computing Center (TACC) for providing raw web log data.

References

[1] B. Schlender. Peter Drucker sets us straight. *Fortune*, (December 29, 2003), http://www.fortune.com.

[2] T.H. Davenport. *Thinking for a Living - How to Get Better Performance and Results from Knowledge Workers*. Harvard Business School Press, Boston, 2005.

[3] A.L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[4] P. Géczy, S. Akaho, N. Izumi, and K. Hasida. Navigation space formalism and exploration of knowledge worker behavior. In G. Kotsis, D. Taniar, E. Pardede, and I.K. Ibrahim, Eds., *Information Integration and Web-based Applications and Services*, pp. 163–172, OCG, Vienna, 2006.

[5] S. AlShaali and U. Varshney. On the usability of mobile commerce. *Intl. J. Mobile Communication*, 3:29–37, 2005.

[6] M. Petre, S. Minocha, and D. Roberts. Usability beyond the website: an empirically-grounded e-commerce evaluation for the total customer experience. *Behaviour and Information Technology*, 25:189–203, 2006.

[7] R. Benbunan-Fich. Using protocol analysis to evaluate the usability of a commercial web site. *Information and Management*, 39:151–163, 2001.

[8] R.J.K. Jacob and K.S. Karn. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyona, R. Radach, and H. Deubel, editors, *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pp. 573–605, Elsevier Science, Amsterdam, 2003.

[9] K.L. Norman and E. Panizzi. Levels of automation and user participation in usability testing. *Interacting with Computers*, 18:246–264, 2006.

[10] J.W. Palmer. Web site usability, design, and performance metrics. *Information Systems Research*, 13:151–167, 2002.

[11] A. Seffah, M. Donyaee, R.B. Kline, and H.K. Padda. Usability measurement and metrics: A consolidated model. *Software Quality Journal*, 14:159–178, 2006.

[12] H. Wu, M. Gordon, K. DeMaagd, and W. Fan. Mining web navigaitons for intelligence. *Decision Support Systems*, 41:574–591, 2006.

[13] P. Géczy, S. Akaho, N. Izumi, and K. Hasida. Extraction and analysis of knowledge worker activities on intranet. In U. Reimer and D. Karagiannis, Eds., *Practical Aspects of Knowledge Management (LNAI)*, pp. 73–85, Springer-Verlag, Heidelberg, 2006.

[14] L. Catledge and J. Pitkow. Characterizing browsing strategies in the world wide web. *Computer Networks and ISDN Systems*, 27:1065–1073, 1995.

[15] M.V. Thakor, W. Borsuk, and M. Kalamas. Hotlists and web browsing behavior–an empirical investigation. *Journal of Business Research*, 57:776–786, 2004.