# COMPLEX USER BEHAVIORAL NETWORKS AT ENTERPRISE INFORMATION SYSTEMS

Peter Géczy, Noriaki Izumi, Shotaro Akaho and Kôiti Hasida
*National Institute of Advanced Industrial Science and Technology (AIST)*

Keywords: Complex networks, web behavior, behavior segmentation, navigation space, knowledge workers, enterprise systems, information services, data mining.

Abstract: We analyze human behavior on a large-scale enterprise information system. Employing a novel framework that efficiently captures complex spatiotemporal dimensions of human dynamics in electronic spaces we present vital findings about knowledge workers' behavior on enterprise intranet portal. Browsing behavior of knowledge workers resembles a complex network with significant concentration on navigational starters. Common browsing strategy utilizes the knowledge of the starting navigation point and recollection of the traversal pathway to the target. Complex traversal network topology has a small number of behavioral hubs concentrating and disseminating the browsing pathways. Human browsing network topology, however, does not match the link topology of the web environment. Knowledge workers generally underutilize the available resources, have focused interests, and exhibit diminutive exploratory behavior.

## 1 INTRODUCTION

Elucidation of human dynamics in electronic environments is of central importance in personalization technologies (Baraglia and Silvestri, 2007), recommender systems (Adomavicius and Tuzhilin, 2005), and collaborative filtering engines (Jin et al., 2006). Corporate sector has been exploring the customer web behavior primarily for commercial purposes (Park and Fader, 2004), (Moe, 2003) and search ranking (Agichtein et al., 2006). Little attention has been devoted to the study of user behavior in enterprise internal information environments. This study presents the scarce results of knowledge worker behavior on a large enterprise intranet portal.

It has been reported that the individual human actions in web environments follow non-Poisson statistics characterized by the long tails (Dezso et al., 2006), (Vazquez et al., 2006). The long tail attributes of human dynamics (Barabasi, 2005) are equivalent to those observed in complex networks (Newman, 2003),(Newman et al., 2005), (Caldarelli, 2007). A common property of complex networks is that the vertex connectivities follow a long tail distribution. The long tiled power-law has been detected in the temporal characteristics of human information access on the web (Dezso et al., 2006). Similar results have been reported from workload studies of search engines and server systems (Bedue et al., 2006),(Schroeder and Harchol-Balter, 2006). The long tails of human interactions have been modeled by power distributions (Vazquez et al., 2006), (Vazquez, 2005), lognormal and Pareto distributions (Downey, 2005), or Zipf distribution (Leskovec et al., 2005).

This work focuses on frequency rather than temporal characteristics of human dynamics in electronic environments, and targets traversal networks of knowledge worker intranet browsing behavior. Applying novel analytic and exploratory framework we present valuable behavioral findings.

## 2 CONCEPT PRESENTATION

User browsing interactions in web environments are reasonably represented by the clickstream sequences. The clickstream sequences of page transitions are segmented into sessions and subsequences. The sessions outline tasks of various complexities, undertaken by the users, that are further divided into the subtasks represented by the subsequences. Segmentation is done according to the users' temporal activity characteristics. Consider the sequence of the form: $\{(p_i,d_i)\}_i$ where $p_i$ denotes the visited page $URL_i$ and $d_i$ denotes a delay between the consecutive views $p_i \rightarrow p_{i+1}$. User browsing activity $\{(p_i,d_i)\}_i$ is



Géczy P., Izumi N., Akaho S. and Hasida K. (2008).
COMPLEX USER BEHAVIORAL NETWORKS AT ENTERPRISE INFORMATION SYSTEMS.
In *Proceedings of the Tenth International Conference on Enterprise Information Systems - HCI*, pages 233-239
DOI: 10.5220/0001700502330239

divided into subelements according to the periods of inactivity $d_i$ satisfying certain criteria.

**Definition 1.** *(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 $B = \{(S_i,ds_i)\}_i$ thus consists of a **train** of subsequences $S_i$ separated by inactivity delays $ds_i$.

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 (Catledge and Pitkow, 1995) indicated that 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 a 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 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 $A \subseteq \mathcal{G}$.

Separation of subspaces within a navigation space reflects 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.

Important aspect to observe in human browsing behavior is to identify the starting and attracting points in navigation space, as well as the 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 refer to the initial navigation points of users, 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) (Thakor et al., 2004).

Page traversal network may contain points that are occasionally accessed and also points concentrating traffic—hubs. Hubs have larger incoming and outgoing spectrum of navigational choices. To quantify a variety of navigational pathways that lead into and out of a point, we define the *in* and *out* degrees.

**Definition 4.** *(In and Out Degrees)*
Let $p_i \in \mathcal{P}$ be a point in a navigation space $\mathcal{G} = (\mathcal{P},\mathcal{B},\mathcal{S})$ such that there exists $B \in \mathcal{B}$ where $|B| > 1$ and $(p_i,d_i) \in B$.
**In degree** of a point $p_i$ is the cardinality of a set of all preceding points $p_{i-1}$ in sessions; $p_{i-1} \rightarrow p_i$, denoted as:
$In(p_i) = |\{p_{i-1}|(p_{i-1},d_{i-1}) \in B \wedge (p_i,d_i) \in B\}|$.
**Out degree** of a point $p_i$ is the cardinality of a set of all following points $p_{i+1}$ in sessions; $p_i \rightarrow p_{i+1}$, denoted as:
$Out(p_i) = |\{p_{i+1}|(p_{i+1},d_{i+1}) \in B \wedge (p_i,d_i) \in B\}|$.

The *in degree* of a point reflects the variety of choices from which the users access it. The point's *out degree* represent the spectrum of branches from it that users utilize. Note that the defined in and out degrees delineate browsing behavior characteristics rather than the number of links pointing to and out of a given point. Some pathways might not be exploited by the users, or users may choose to utilize hotlists at a given browsing stage. The human browsing behavior hubs in the navigation space may differ from the link hubs.

# 3 INFORMATION SYSTEM CASE STUDY

The information system investigated in this study is the large-scale intranet portal of The National Institute of Advanced Industrial Science and Technology. The core comprises of six servers connected to the high-speed backbone in a load balanced configuration. The accessibility is provided via wide ranging connectivity options (from high-speed optical to wireless) accommodating several platforms (up to mobile devices). The portal provides extensive range of web services and documents vital to the organization (Table 1). The rich intranet services support business processes for management, accounting and administration, research cooperation with industry and other institutes, and resource localization; but also bulletin boards and networking within organization. The institute has a number of branches throughout the country, thus several services and resources are distributed. Visible web space exceeded 1 GB, and deep web space was substantially larger, but difficult to estimate due to the decentralized architecture and varying back-end data.

Table 1: Case study data information.

| | |
|---|---|
| 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 |
| Log Records | 315 005 952 |
| Clean Log Records | 126 483 295 |
| Unique IP Addresses | 22 077 |
| Services | 855 |
| Unique URLs | 3 015 848 |
| Scripts | 2 855 549 |
| HTML Documents | 35 532 |
| PDF Documents | 33 305 |
| DOC Documents | 4 385 |
| Others | 87 077 |
| Sessions | 3 454 243 |
| Unique Sessions | 2 704 067 |
| Subsequences | 7 335 577 |
| Unique Subsequences | 3 547 170 |
| Valid Subsequences | 3 156 310 |
| Unique Valid Subsequences | 1 644 848 |
| Users | ∼10 000 |

The majority of the enterprise portal users were skilled knowledge workers. Significant traffic on the portal resulted in a large web log data pool. The traffic was both human and machine generated, thus the data required cleaning. The data preparation, processing, filtering, and segmentation to sessions and subsequences are described in (Géczy et al., 2007). The initial data cleaning eliminated most of the machine generated traffic, however, further filtering was needed after subsequence extraction. It is noticeable that the data cleaning and filtering reduced the number of log records by 59.85%, as well as the number of unique valid subsequences by 53.6%.

# 4 BROWSING BEHAVIOR ANALYSIS

By analyzing the point characteristics we infer several relevant observations. The point characteristics of a navigation space highlight the initial and the terminal targets of knowledge worker activities, and also the single-action behaviors. Analysis demonstrates the applicability and usefulness of the approach.

It is evident that knowledge worker navigation space is substantially smaller, with respect to the essential navigation points, than the observed complete navigation space. The unique valid sets of starters (115770), attractors (288075), and singletons (57 894) are very small in comparison to the set of unique URLs (3015848) in the navigation space (see Table 1 and Table 2). The largest set, unique valid attractors, is only 9.55% of unique URLs. Unique valid starters and singletons represent only approximately 3.84% and 1.92% of unique URLs, respectively.

*Browsing behavior of knowledge workers resembles the complex networks.* Topology of knowledge worker navigation space clearly corresponds to the complex network. Characteristic feature of complex networks is a long tailed distribution of the in and out degrees of the nodes. Histograms of in and out degrees of starters and attractors distinctly display long tail characteristics–with small number of high frequency elements gradually progressing to the large number of low frequency elements (Figure 1 and 2). The network of starting navigation points as well as the network of users' targets are both complex networks. Certain points in the navigation space concentrate the human web traffic and serve as hubs.

*Knowledge workers' browsing behavior concentrates on the navigational starters.* Starters are the major concentration points of the users' complex navigational network. They are the main hubs. There are approximately one hundred primary starter hubs and three hundred primary attractor hubs. These one

Table 2: 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 |

hundred primary starters constitute 0.086% of unique valid starters, and three hundred primary attractors account for 0.1% of unique valid attractors. Thus the ratio between the primary starter and attractor hubs is approximately one to three. This one-to-three ratio approximately holds also between the numbers of unique valid starters (115 770) and attractors (288 075) — see Table 2.

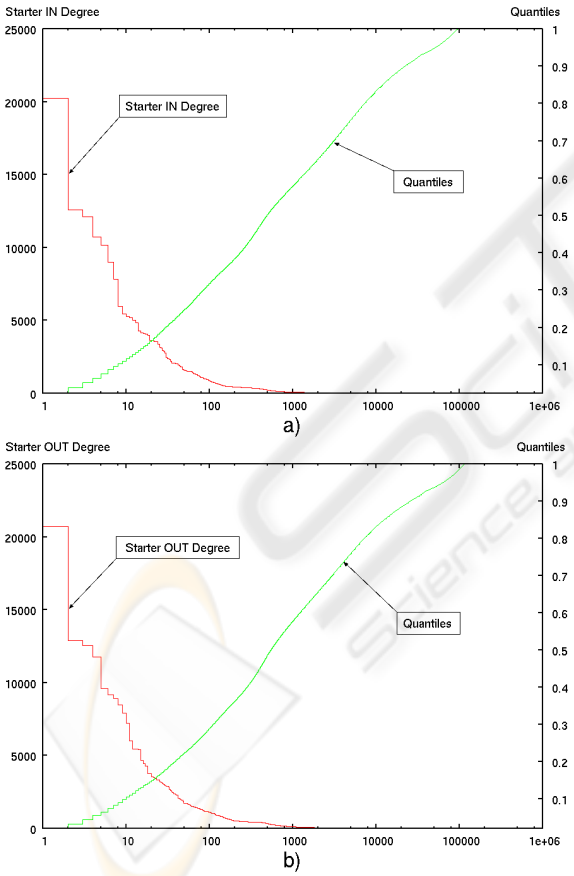


Figure 1: Histograms and quantiles of starter: *a)* in degrees, *b)* out degrees. Right y-axis contains a quantile scale. X-axis is in a logarithmic scale.

*The initial navigation points primarily disseminate the knowledge worker browsing pathways.* The starters disperse the navigation more than the attractors. This is evident from the quantification of the in and out degrees of the major starters and attractors. In and out degrees of starters range from one to over twenty thousand. Range of attractor in degrees (1 to about 6800) and out degrees (1 to about 3400) is approximately three to six times lower, respectively. Top ten starters (approximately 0.0086% of unique valid starters) have in and out degrees ranging from five thousand to over twenty thousand (Figure 1). Compound in and out degrees of top thirty starter hubs (approximately 0.026% of unique valid starters) represented approximately 20% of total starter in and out degrees.

*Knowledge workers are more behaviorally diverse in reaching their targets than proceeding to the starting points of the following sub-tasks.* The attractors' in degree range is two times greater than the out degree range (refer to Figure 2). Thus the users employ approximately two times more arriving pathways to the targets than the departing ones. They are more diverse in reaching the targets than proceeding to the following navigation points of the consequent subtasks. Only approximately top twenty attractors have in and out degrees greater than one thousand. Discrepancies between their in degrees are greater than between their out degrees.

Variability of arriving and departing pathways to and from starters is relatively balanced. Both, in and out degrees of starters extend to approximately 20000 (Figure 1). The in and out degree ranges of starters are significantly greater than the attractor ranges (see Figures 1 and 2). Hence the users have richer traversal repertoire when reaching and leaving the initial navigation points rather than the targets.

*Knowledge workers utilized a small spectrum of starting navigation points and targeted relatively small number of resources during their browsing.* The set of unique valid starters (115770), i.e. the initial navigation points of knowledge workers' (sub-)goals, was approximately 3.84% of total navigation points (see Tables 1 and 2). Although the set of unique valid attractors (288075), i.e. (sub-)goal targets, was approximately three times higher than the set of initial navigation points, it is still relatively minor portion

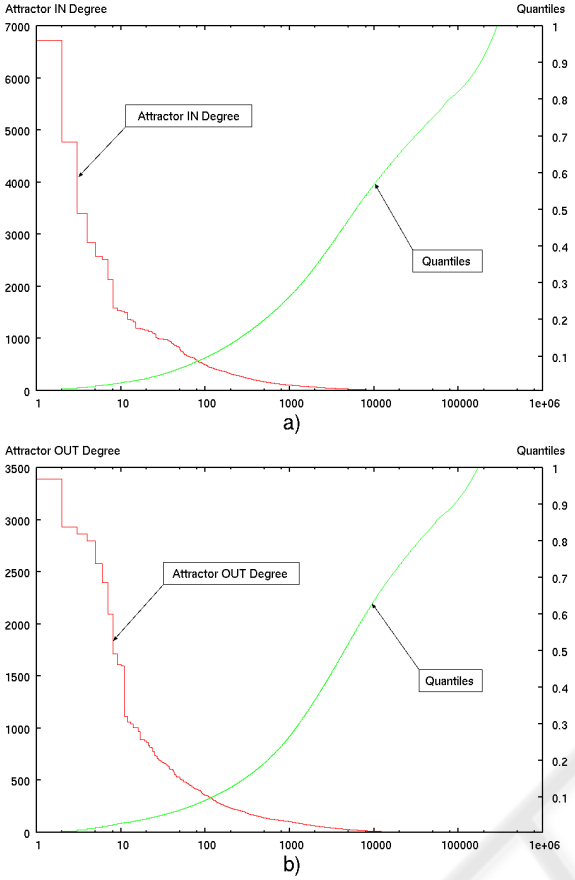


Figure 2: Histograms and quantiles of attractor: *a)* in degrees, *b)* out degrees. Right y-axis contains a quantile scale. X-axis is in a logarithmic scale.

(approximately 9.55% of unique URLs). Knowledge workers initiated their browsing experiences from a small number of navigation points and aimed at relatively few resources.

*Few resources were perceived of value to be bookmarked.* Number of unique single user actions was minuscule. Single actions, such as use of hotlists (Thakor et al., 2004), followed by delays greater than 1 hour are represented by the singletons. Unique valid singletons (57894) accounted for only 1.92% of navigation points (see Tables 1 and 2). The number of singletons is approximately two times lower than the number of starters and almost five times lower than the number of attractors (Table 2). If only small number of starters and/or attractors were perceived useful, there is a possibility that they were bookmarked and accessed directly in the future browsing experiences.

*Knowledge workers had focused interests and exhibited minuscule exploratory behavior.* A narrow spectrum of starters, attractors, and singletons was frequently used. The histograms and quantile characteristics of starters, attractors, and singletons (see

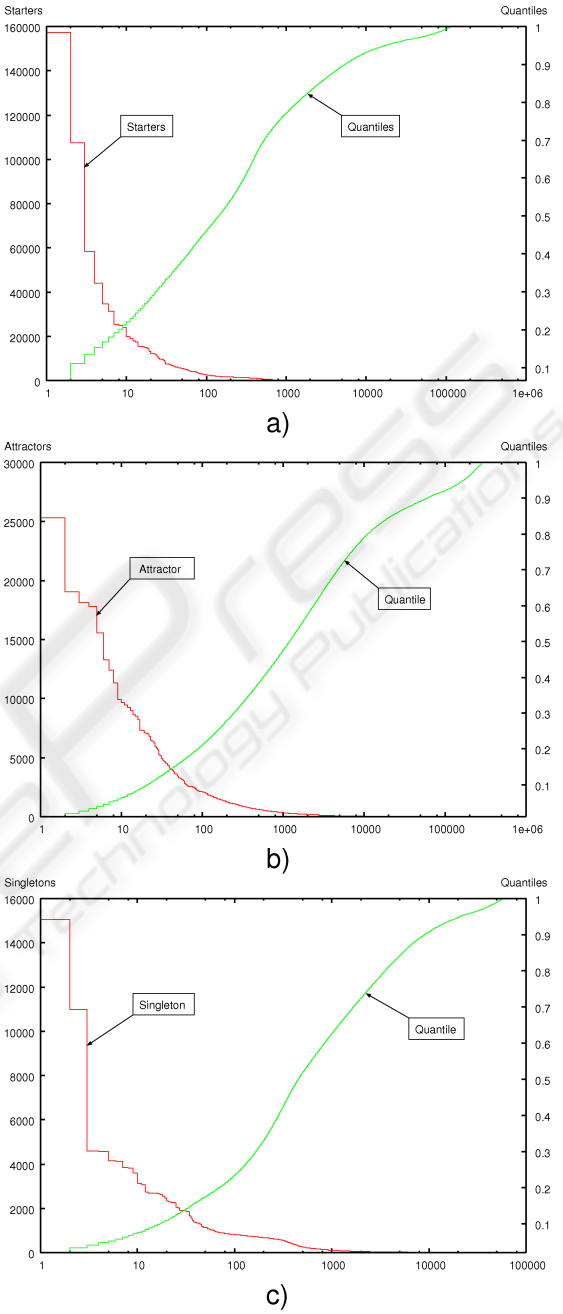


Figure 3: Histograms and quantiles: *a)* starters, *b)* attractors, and *c)* singletons. Right y-axis contains a quantile scale. X-axis is in a logarithmic scale.

Figure 3) indicate that higher frequency of occurrences is concentrated to relatively small number of elements. Approximately ten starters and singletons, and fifty attractors were very frequent. About one hundred starters and singletons, and one thousand attractors were relatively frequent. The quantile analysis in Figure 3 reveals that ten starters (0.0086%

of unique valid starters) and singletons (0.017% of unique valid singletons), and fifty frequent attractors (0.017% of unique valid attractors) accounted for about 20% of total occurrences. One hundred starters (0.086% of unique valid starters) and one thousand attractors (0.35% of unique valid attractors) constituted about 45% and 48% of total occurrences, respectively. Analogously, one hundred twenty singletons (0.21% of unique valid singletons) compounded to about 37% of total occurrences.

*Knowledge workers were generally more familiar with the starting navigation points rather than the targets.* Smaller number of starters repeats substantially more frequently than the adequate number of attractors. That is, the users knew where to start and were familiar with the navigational path to the target (instead of just utilizing shortcuts such as bookmarks). In and out degrees of frequent starters are also significantly higher than those of attractors (see Figures 1 and 2). The frequent starters have in and out degrees between 5000 and 20000, whereas the frequent attractor in degrees are between 1000 and 6800, and out degrees between 1000 and 3400.

*Complex networks of knowledge worker browsing behavior differ from the web topology constituted by links.* Hubs in the web topology are the pages with large number of incoming and outgoing links. Behavioral hubs are the navigation points that have large in and out degrees — resulting from the user traversal patterns. It has been discovered that the behavioral hubs in the knowledge worker navigation space did not substantially match the link hubs. High out degrees of behavioral hubs (reaching almost 7000) also significantly exceed the number of links on the served pages at any given time.

## 5 CONCLUSIONS

We introduced a novel analytic framework for exploration and modeling of human browsing behavior in electronic environments. It utilizes a temporal segmentation of browsing activity. The framework was applied to browsing behavior analysis of the knowledge workers on a large enterprise information system. Numerous vital behavioral features have been revealed. Knowledge worker browsing behavior concentrated on the navigational starters. They remembered the starting point and recalled the navigational path to the target. The knowledge workers effectively utilized only a small amount of available resources. A large number of resources have been occasionally accessed.

Topology of knowledge worker traversal pathways resembles complex networks. However, the behavioral complex network differs from the hypertext link network. The traversal hubs do not identically correspond to the link hubs. Significant long tail characteristics of the essential navigation points have been exposed both in terms of frequencies as well as in and out degrees.

## REFERENCES


Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17:734–749.

Agichtein, E., Brill, E., and Dumais, S. (2006). Improving web search ranking by incorporating user behavior information. In *Proceedings of The 29th SIGIR*, pp. 19–26, Seattle, Washington, USA.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211.

Baraglia, R. and Silvestri, F. (2007). Dynamic personalization of web sites without user intervention. *Communications of the ACM*, 50:63–67.

Bedue, C., Baeza-Yates, R., Ribeiro-Neto, B., Ziviani, A., and Ziviani, N. (2006). Modeling performance-driven workload characterization of web search systems. In *Proceedings of CIKM*, pp. 842–843, Arlington, USA.

Caldarelli, G. (2007). *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford University Press, Cambridge, UK.

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

Dezso, Z., Almaas, E., Lukacs, A., Racz, B., Szakadat, I., and Barabasi, A.-L. (2006). Dynamics of information access on the web. *Physical Review*, E73:066132(6).

Downey, A. (2005). Lognormal and pareto distributions in the internet. *Computer Communications*, 28:790–801.

Géczy, P., Akaho, S., Izumi, N., and Hasida, K. (2007). Usability analysis framework based on behavioral segmentation. In Psaila, G. and Wagner, R., Eds., *Electronic Commerce and Web Technologies*, pp. 35–45, Springer-Verlag, Heidelberg.

Jin, R., Si, L., and Zhai, C. (2006). A study of mixture models for collaborative filtering. *Information Retrieval*, 9:357–382.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of KDD*, pp. 177–187, Chicago, Illinois, USA.

Moe, W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13:29–39.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Newman, M., Barabasi, A.-L., and Watts, D. (2005). *The Structure and Dynamics of Complex Networks*. Princeton University Press, Princeton, N.J.

Park, Y.-H. and Fader, P. (2004). Modeling browsing behavior at multiple websites. *Marketing Science*, 23:280–303.

Schroeder, B. and Harchol-Balter, M. (2006). Web servers under overload: How scheduling can help. *ACM Transactions on Internet Technology*, 6:20–52.

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

Vazquez, A. (2005). Exact results for the barabasi model of human dynamics. *Physical Review Letters*, 95:248701(6).

Vazquez, A., Oliveira, J., Dezso, Z., Goh, K.-I., Kondor, I., and Barabasi, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review*, E73:036127(19).