A longitudinal study of World Wide Web users' information searching behavior

Vivian Cothey

Graduate School of Education, University of Bristol

35 Berkeley Square, Bristol, BS8 1JA, United Kingdom

telephone: (44) 117 928 7105, fax: (44) 0117 925 1537

email viv.cothey@bris.ac.uk

A study of the "real world" Web information searching behavior of 206 college students over a ten month period showed that, contrary to expectations, the users adopted a more passive or browsing approach to Web information searching and became more eclectic in their selection of Web hosts as they gained experience. The study used a longitudinal transaction log analysis of the URLs accessed during 5431 user-days of Web information searching to detect changes in information searching behavior associated with increased experience of using the Web. The findings have implications for the design of future Web information retrieval tools.

Introduction

Investigators have long attempted to discover relationships between a user's information searching behavior when using electronic information systems and the user's "expertise", that is, whether the user is a *novice* or *expert* (for example, Fenichel, 1981; Harris, 1986; Hsieh-Yee, 1993; Lazonder et al., 2000). Hence an objective of research is to discover differences between the information searching behavior of novices and experts that distinguish them. Defining expertise and distinguishing novices from experts is difficult and as a consequence *experience* is generally substituted for expertise. It is assumed, either implicitly or explicitly, that more experienced users are more expert (Fidel et al., 1998; Khan & Locatis, 1998). In addition it is generally assumed that more expert users will use more *systematic* search strategies (Iivonen, 1995; Marchionini et al., 1993; Sutcliffe et al., 2000). Like expert, a precise definition of what is systematic is also problematic. However, ever since Cove & Walsh (1988) showed that users' *searching* and *browsing* strategies could be distinguished, investigations have classified user searching behavior within a broad spectrum having "searching" at one end and "browsing" at the other (for example, Catledge & Pitkow, 1995; Qiu, 1993). Various criteria have been used to do this which have usually been subjective and have appealed to the notion that searching is more *analytic* than browsing (Carmel et al., 1992; Chen et al., 1998; Schater et al., 1998). Marchionini (1995, p.8) suggests a fundamental distinction between searching (or analytic) search strategies and browsing search strategies:

> *Analytical* strategies depend on careful planning, the recall of query terms, and iterative query reformulations and examinations of results. *Browsing* strategies are heuristic and depend on recognizing relevant information.

Analytic (or "searching") information searching strategies require a user to be more active than does a "browsing" strategy when the user is more passive.

Although there are many novice/expert investigations including a recent study of Web users (Hölscher & Strube, 2000), "longitudinal" studies of information retrieval behavior are rare (Yuan, 1997) and few longitudinal studies of Web user searching strategies have been reported. Previous novice/expert investigations are "cross-sectional", that is they compare two groups of users at essentially the same point in time, whereas in this investigation we use a longitudinal-developmental research design based on repeated measures relating to the same individual (Nesselroade & Baltes, 1979). In this investigation we make use of transaction log data that records over a ten month period all the uniform resource locators (URLs) that are accessed by each of 206 student users. This data is analyzed in order to detect changes in a user's Web information searching behavior as that user gains in experience of using the Web. Hence this study monitors the same user at different points in time so that each user studied is gaining experience during the period of the investigation. This is a typical characteristic of longitudinal-developmental investigations (Goldstein, 1979).

The study took place in the *real world* (Jansen et al., 2000) context of the user's Web information seeking so that the users' information tasks were not imposed for the purposes of the investigation, rather the information tasks are self-constructed by the users and reflect the users' own particular information needs (Gross, 1999; Spink et al., 1998).

There is little underlying theory of Web information searching as distinct from information searching theory more generally and especially information searching in electronic environments (Marchionini, 1995). Pirolli & Card (1999) have developed an ecological approach based on foraging theory (Bell, 1990) however this is restricted to modeling the movements of groups of users within and between different Web sites (Huberman et al., 1998, 3 April). Understanding users' access to the broad range of Web sites available to them is an open theoretical question.

Related studies

Catledge & Pitkow (1995) and Cunha et al. (1995) undertook the first cross-sectional client-side Web user investigations into URLs accessed (Pitkow, 1998). These studies made use of modifications to the Mosaic Web browser software, then in common use, in order to compile a transaction log. This form of instrumentation can have the advantage of collecting data about other user events such as use of the back-button which Tauscher & Greenberg (1997) exploited in their human-computer interaction study of Web usage. Although current proprietary browser software cannot be specially instrumented, additional Web user monitoring software (for example, Optimal, Graham-Cumming, 1997) can be attached to

local networks or installed in user's machines at home (for example, HomeNet, Kraut et al., 1996). URL requests can also be intercepted by a proxy server and the server log file used to provide a transaction log. Choo et al. (1998) used both techniques to study eleven users over a two week period. Like the previous cross-sectional studies, the users were IT professionals. Abrahamson (1998) reports a larger investigation based on using the Optimal software to monitor Web usage from publically available computers during four, one week sample periods. This investigation is potentially attractive because of the "public" user population but the study provides no further information about the users.

Longitudinal Web studies are reported by Christ et al. (2001), Cockburn & McKenzie (2001) and Cooper (2001). Cooper investigates changes during a 479 day period in the usage access patterns to a single Web site, a library catalog and recommends future examination of individual user behavior. Both Christ et al. and Cockburn & McKenzie adopt a longitudinal-developmental approach in that they examine the changes in behavior of individual Web users.

Christ et al. examined use of the Web at home (the HomeNet project) and analyzed transaction logs from up to 339 individuals collected over a 156 week period. Their analysis focuses on the change in the number of different Web sites which an individual visits each week. They found that as individuals become more experienced so the number of different Web sites visited reduced to a "saturation" level which depended upon whether the individual was a "moderate, heavy or very heavy" Web user.

Cockburn & McKenzie specifically identified that the empirical work by Catledge & Pitkow needed to be brought up to date because of the growth of the Web. They therefore investigated the URLs accessed by 17 individual users over a period of 119 days. In addition to improving the currency of the empirical data, Cockburn & McKenzie also aimed to overcome particular deficiencies in the earlier work such as users not able to use their preferred Web browser and the sample period lasting only 3 weeks. Since all Cockburn & McKenzie's users now used Netscape's Navigator Web browser and were part of a managed local network they were able to collect daily copies of each individual's Navigator `history.dat` file as part of the institution's normal incremental back-up procedures. This file contains details of each URL accessed by the individual, including URLs retrieved from the browser's internal cache memory that are not recorded by external software monitors. Cockburn & McKenzie describe (pp.906–908) how they compiled transaction logs for each individual from their `history.dat` files and note that these are Berkeley DB 1.85 hash files (Olson et al., 1999) not straightforward text files. Like Christ et al., Cockburn & McKenzie focus their analysis on the changes in the user's *vocabulary*, that is the number of different Web sites that an individual visits, and an individual's revisitation rate, that is visiting a Web site that the individual has already visited. They report that, although compared to earlier studies users

are visiting more Web sites, “there is a marked lack of commonality in the sites visited by different users” (Cockburn & McKenzie, 2001, p.903). A possible limitation of their study is that, once again, the users were IT professionals.

Research aim

The aim of this investigation is to detect whether or not there is any change in an individual’s Web information seeking behavior as that individual gains experience. In particular two aspects of change in an individual’s behavior as the individual gains experience are hypothesized that would be expected if more experienced users became more systematic in their information seeking. These are;

1. that individuals would increase their active information seeking, for example by an increased use of search engines, and
2. that there would be increasing commonality among individuals in the selection of Web sites as more useful sites are revisited and less useful sites ignored.

Research design and methodology

Overview

The research design is a large scale cohort based longitudinal-development study using Web access transaction logging where repeated measures are taken in respect of each individual. Anonymous recoding of unique user identification ensures user privacy yet allows individual transaction logs to be compiled that record the URLs requested by each of 206 users over a period of ten months. The user sample is a cohort drawn from the student population at a Higher Education institution in the UK. The URLs from each user’s Web information seeking are divided into daily *sessions* and analyzed to compute two metrics, *user querying rate* and *Web host conformance* as described below. A vector model approach is used to measure the similarity/dissimilarity of the set of Web sites visited by an individual when compared to the cohort. The conditional regression model of longitudinal analysis (Plewis, 1985) is used to detect change over time.

The user sample

The user sample is drawn from the general population of students at a Higher Education institution in the UK rather than from a single academic department such as computer science or information science. Unrestricted user access to the Web is freely available following logging on to the institution's local network and is provided via the institution's high bandwidth connection to the UK's academic network. The user sample from the transaction log consists of the cohort of all those Web users who first registered at the start of the 1997/8 academic year (that is incoming freshmen) and who attend a full-time course of study. These criteria therefore exclude students attending part-time (which include a variety of post-professional qualification short courses) and "sandwich" type courses involving work placements. Hence, the sample of 206 users is homogeneous (compared to the student population as a whole) in the sense that the sample is a single cohort of students all having similar potential previous Web usage experience and all sharing the characteristic of being full-time students. The gender split in the sample is 98 men and 108 women.

Data collection

The data used for this study is a Web transaction log. Transaction log data typically consists of an electronic record of the user's information searching request of an information system, or query, and the system's response. These data are often complemented by a *timestamp* which records when the request/response occurred. Web transaction logs generally provide a record of the URL accessed together with a timestamp.

The Web transaction log data for this investigation were collected by copying the daily individual *history* files automatically generated by the Netscape Navigator browser then in use. Each individual user has a unique cumulative history file maintained on a central server accessible only in response to a valid user login. In consequence the data could be associated with a particular user and a complete user transaction log covering the 1998-1999 academic year (October 1998 to June 1999) was compiled for each individual. This data collection method is equivalent to that used by Cockburn & McKenzie as noted above.

In order to ensure user privacy, identifiable user codes were recoded to produce consistent but anonymous individual user codes (Penniman & Dominick, 1980).

A key concept embodied within the transaction monitoring technique is that of an information searching *session* (He & Göker, 2000), that is, a demarcating of the collection of transaction records so that together

they constitute (for a particular user) a self contained episode of information searching (Borgman et al., 1996; Peters, 1993). This study takes a pragmatic approach to defining a user's Web session in that all the Web transaction records generated by a user during a single day constitute a session rather than either the fixed time threshold most commonly used (Pitkow, 1997) or the "different host" approach by Huberman et al. (1998, 3 April).

The daily history file corresponds naturally to the daily information searching session of the Web transaction log analysis, and, since it is a *client-side* record of the Web transaction it includes all the URLs accessed (both explicitly and implicitly) by the user. This property of client-side data contrasts with Web *server-side* data which generally relates only to URLs accessed at a single Web host. Server logs can provide detailed information about the population of users accessing a particular Web host but generally fail to say anything about the population of Web host accesses from a particular user. An exception to this is the cross-sectional study by Huberman et al.

Compared to other client-side data collection techniques (such as video recording or software monitoring), transaction log analysis using the history files is unobtrusive, does not degrade the performance of the browser software (which could affect the user's behavior) and can be used on a large scale both in respect of the numbers of users studied and the length of the study period. However the transaction log analysis technique is intrinsically limited (Kurth, 1993) and the history file in particular does not reveal low level human-computer interaction events.

Data analysis

The investigation's data analysis combines the technique of Web transaction log analysis with the conditional regression model of longitudinal analysis (Plewis, 1985).

The longitudinal analyses uses a split-half technique. That is the collection of daily sessions for each user is considered as a sequence in chronological order and split into two chronological halves, an earlier and later half (King, 1991, p.365). The information searching behavior during the earlier half is compared with searching behavior during the later half in order to investigate the change phenomenon associated with the user's increasing experience. For example, a user who had recorded sessions on days 2, 7, 12, 22, 50 and 59 of the study period would have an earlier half period of sessions 2, 7 and 12 and a later half period of sessions 22, 50 and 59. If there was an odd number of sessions involved then the extra session was included in the earlier half. The half period split therefore divides a user's Web information searching experience into two similar portions in a way which corresponds to the real world of the users'

Web information searching behavior rather than to any arbitrary calendar based method. Clearly, in the absence of any change then the later half information searching behavior will be (statistically) the same as the earlier half information searching behavior.

A modification of the standard term-document vector model is used to compute the user's Web host conformance metric, see below. The paradigm of document collection, document, and term is used throughout. Hence the methodology of a Web transaction log analysis can equally be regarded as a document collection analysis. That is, each user's daily Web information searching session is represented as a *document* containing a set of *terms* i.e. each URL accessed during the session. The analysis filtered out URLs that were accessed implicitly by a user such as those URLs that refer to graphical image files included in a Web page being explicitly requested. The standard term-document vector is therefore an n-tuple having entries being the frequency of access to a URL requested by the user during the session.

The Web transaction log analysis is also an analysis of URLs. The string of characters forming each URL, for example;

`http://www.bristol.ac.uk/Depts/Info-Office/about/facts.htm`

contains the component parts;

network host name which here is given by `www.bristol.ac.uk`, and

path which here is given by `Depts/Info-Office/about/facts.htm`

Optionally the URL path may conclude with a *searchpart* which is by definition preceded by a `?` (Berners-Lee et al., 1994). This occurs when the user is querying.

The *host part* may be in *domain name* form as above, or it may be in *dotted decimal address* format such as 137.222.10.46. Domain names are composed of a restricted character set and, in particular, all uppercase characters are interpreted as lowercase. Upper and lowercase character distinctions are allowed in URL paths. Hence although valid URLs have a canonical format (Berners-Lee et al., 1994) the same Web host name may have several representations. Web host names or representations were therefore normalized by making use of the Domain Name Service (DNS) *official name* (Mockapetris, 1987). All Web hosts are accessed via a machine readable Internet Protocol (IP) address. This IP address (or addresses - there may be more than one) can be represented legibly as a dotted decimal address while most Web hosts have one or more domain names such as `www.bristol.ac.uk`. The DNS within the Internet translates from domain names to IP addresses.

The following algorithm normalizes the domain names or terms to produce a unique standard representation;

1. the host part of the URL is extracted and validated as being in canonical form
2. any uppercase characters in the host part are converted to lowercase
3. the default term is set equal to the host part
4. the DNS entry for each host part is requested and if this is available then the DNS official name replaces the default term
5. finally all leading "`www.`" characters on terms are removed to leave the normalized term version of host part of the URL.

Hence valid domain name variants of URLs which all access the identical Web host are given a single standard representation. For example, `www.Bristol.ac.uk`, `bris.ac.uk` and `137.22.10.46` are normalized to `fsa.bris.ac.uk`.

The data analysis uses two specific metrics to quantify an individual's Web searching/browsing behavior and which correspond to the research hypotheses stated under **Research aim**. These are;

1. *user querying rate*, and
2. *Web host conformance*.

The construction and rationale of these metrics are described in detail below. As noted, vocabulary based metrics have also been used in longitudinal Web user studies and several studies have examined in detail the searchpart of Web transaction logs (for example, Spink & Xu, 2000). Similar metrics will be used in future analyses of this data.

User Web information searching metrics: user querying rate

The user querying rate aims to quantify a user's active information seeking. There are two styles of Web information seeking, "querying" and "link-clicking". When querying, a user submits particular data to a Web server which is used to determine a response while when link-clicking, a user passively (Palmquist & Kim, 2000) follows simple hypertext presented on the Web page being viewed in order to request access to the next Web page. Hence, in the context of Web information searching, querying

corresponds to Marchionini's (1995) more active analytical strategies while link-clicking corresponds to his more passive browsing strategies. When querying, the user is being active in the sense that the user is submitting particular information to a Web host but when link-clicking a user is more passive in that no user-particular information is generated. Querying may be highly specific as in submitting a query to a Web search engine, or it may be implicit as can occur when the user selects from some automated choice criteria presented on the Web page being viewed. Since the transaction log for each user's Web information searching session contains all the different URLs submitted by the user an analysis can determine categorically whether the user has queried or link-clicked. This is because in the former case the URL submitted contains a "?". This is absent in the latter (Berners-Lee et al., 1994).

The content of the query, that is the searchpart, is ignored for the purpose of the analysis and no distinction is drawn between a user submitting just a single query during a (daily) session and a user submitting multiple queries. If any of the daily URLs requested indicates querying by the user then the session as a whole is classed as a querying session. The user's querying rate is defined to be the proportion of the user's sessions which are querying sessions.

User Web information searching: Web host conformance

The Web host conformance measure aims to quantify the commonality among users in their selection of Web sites. That is, the host conformance is a similarity/dissimilarity measure (Boyce et al., 1994) for the collection of Web hosts accessed by a particular user when compared to the overall collection of Web hosts accessed by all users in the study. The value of the Web host conformance measure in a particular instance is computed using a modification (as described below) of the standard term-document vector model (Salton & McGill, 1983). To apply this approach, each user's Web information seeking session is regarded as a "document" containing a set of "terms" where each term is the normalized host name of a Web host accessed by the user during the session.

Each session can be represented by its corresponding binary term vector. As usual, this vector is transformed into a scalar value by forming its inner vector product with another vector under a term weighting scheme. For each session (or collection of sessions) this procedure produces the Web host conformance measure which is a real number between 0 and 2. The closer the match between the Web hosts appearing in a particular user's session and the most frequently selected Web hosts overall then the larger will be the user's Web host conformance measure. Conversely the more eclectic the user is in choice of Web hosts then the smaller will be the conformance measure.

The term-document vector model is usually applied to the problem of determining the similarity/dissimilarity between a given user query and individual documents in a collection. Both the query and the documents are expressed as term vectors and the similarity between the query and each document determined (for example, by a procedure involving computing the cosine of the angle between their vectors). An extension of this application is to use the term-document vector model to obtain clusters of either documents from a collection or of the queries submitted by users of an information retrieval system. In each instance the motivation is to obtain a better insight into useful ways to discriminate between the documents or queries. Frants et al. (1993) observed that one could equally well use the vector model procedure to determine a user profile based on the set of queries submitted by a user. Hence, instead of using document terms in the form of a term-document vector to represent a document, the terms of queries submitted by a user are taken to represent the user. Fu et al. (1999) attempts a similar user focused analysis but like Huberman et al. (1998, 3 April) substitutes Web host based sessions for users.

For this investigation each user is represented by a sub-collection of "documents" (user sessions) and each "document" is represented as a term-document vector. The goal of the study is to determine similarities/dissimilarities among these sub-collections.

Both a mean Web conformance and an aggregate Web conformance are computed. The mean Web conformance is the arithmetic mean of the conformance metric calculated for each of an individual's "documents" in either the earlier or later split half-period. The aggregate Web conformance considers the cluster of documents in either the earlier or later period as a single entity. As a consequence the aggregate Web conformance of a user omitting to access a popular Web host during one session is compensated by that user including the Web host during another session within the same half period.

Modified term-document vector model

The generic term-document vector model is usually implemented using a term weighting scheme based on term frequencies and inverse document frequencies, (tf*idf), and a similarity or distance function based on the cosine of the angle between the query vector and the document vector. Documents sufficiently similar or close to the query are retrieved. In the hybrid or extended term-document vector model (that is, embracing more of the Boolean model) the term frequencies are taken to be either 0 or 1 (for example, Salton et al., 1983).

The essential ingredients of the vector model are;

document vectors - which analyze the term composition of each document in a collection,

term weighting - a scheme which biases the similarity/dissimilarity function in favor of terms deemed to be more important, and

similarity/dissimilarity function - a mathematical procedure which operates on a weighted document vector in order to produce a number.

Although there is a consensus in support of using a logarithmic inverse document frequency based weighting scheme and a cosine distance function for query based relative relevance ranked document retrieval systems, the choice of a particular weighting scheme and distance function is pragmatically dependent on the detailed requirements of the application (Salton & Buckley, 1988). Document ranking is here described as relative because although all documents in the collection can be included in the ranking, rank position is relative to the arbitrary query: a different query would produce different rankings.

For this investigation we wish to discriminate between all the documents (and between clusters of documents) in the collection in a way which permits identification of differences in the occurrence of the principal terms in documents. That is, the entire collection should be ranked. A secondary consideration is that the procedure should be as computationally straightforward as possible. Therefore the binary term vector for each document in the collection is transformed via an exponential weighting scheme and similarity function to a number θ between 0 and 2 having in particular the property that the more important are the terms by which the documents differ, the greater will be difference in the two θs. If θ is the same for two documents then the documents contain the same principal terms. The procedure also fully discriminates between different documents in that sufficient terms are considered to always distinguish between different documents.

The investigation's use of the hybrid Boolean tf*idf term-document vector model is referred to as a modified term-document vector model because uses term frequency *ranks* rather term frequencies (Aalbersberg, 1994).

Each user session, D, is represented by a binary document vector $\boldsymbol{d} = (tb_1, tb_2, tb_3, \ldots tb_T)$ where tb_i is a binary variable indicating the occurrence of the normalized term t_i in the document D. The document frequencies, $df(t_i)$, are straightforward to compute since, summing across the document collection,

$$df(t_i) = \sum_{D_1}^{D_N} tb_i$$

Without loss of generality, the components of $\boldsymbol{d}$ can therefore be taken to be in descending rank order

where $df(t_i) > df(t_{i+1})$ so that $rank(t_i) = i - 1$, starting at rank 0 for first place. (The effect of ties is not material.)

Now define the term weighting scheme

$$w_{i,j} = tb(t_{i,j}) \frac{1}{2^{rank(t_i)}} = tb(t_{i,j}) \left(\frac{1}{2}\right)^{i-1}$$

where $tb(t_{i,j})$ is the binary variable indicating the occurrence of term t_i in document D_j. The similarity function $sim(Q, D)$ where $\boldsymbol{q}$ is the binary vector $\boldsymbol{q} = (qb_1, qb_2, qb_3, \ldots qb_T)$ with qb_i being binary variables, is given by

$$sim(Q, D_j) = \sum_{i=1}^{T} qb_i.w_{i,j} = \sum_{i=1}^{T} qb_i.tb(t_{i,j}) \left(\frac{1}{2}\right)^{i-1}$$

Hence if we set $qb_i = 1$ for all i, that is Q is the "unit document" indicated by $Q = 1$, then

$$sim(1, D_j) = \sum_{i=1}^{T} tb(t_{i,j}) \left(\frac{1}{2}\right)^{i-1}$$

and $sim(1, D_j) \mapsto \theta$ maps the binary document vector representing D_j to a number which is the sum of the first T terms of the geometric sequence $\left(\frac{1}{2}\right)^{r-1}$, that is 1, $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16} \ldots$ but with some terms missing.

In practice only the first 30 normalized ranked terms were considered since these provided a *discriminating* set of terms small enough to fall within the computational arithmetic used. If a smaller set of terms were used then at least two documents would be erroneously considered identical.

In the context of this investigation, since the document terms are hosts then saying that two documents have close θs, that is there is a small numerical difference, corresponds to the users accessing nearly the same collection of hosts, that is differing by only less popular hosts. If the collection of hosts accessed by two users differ in respect of access to a very popular host then the corresponding θs will have a large numerical difference. For example, the absence of the most popular discriminating term from a document will reduce its Web host conformance by 1 whereas the absence of the least discriminating term will reduce its Web host conformance by only $\left(\frac{1}{2}\right)^{29}$. Hence θ provides a measure of the extent to which the user conforms to the overall host popularity when accessing hosts for Web information searching. That is θ measures the level of commonality of selection of Web hosts among individuals.

Various rank correlation statistics (Siegel & Castellan, 1988) are available but would be difficult to use

because the data is unbalanced, that is some documents contain only a few terms while others contain many terms.

The conditional regression model

The longitudinal-developmental research design provides repeated measures on the same individual throughout the period of study. However in this user focussed investigation no consistency is expected regarding either the total number of measurements taken per user or the time intervals between the measurements.

In these circumstances conditional regression offers a standard statistical technique for analyzing the phenomenon of change over time in an individual (Plewis, 1985). The technique relies on having a pair of measurements for each individual, say for example, a user's Web querying rate - the proportion of the user's daily Web information searching sessions that have included the user per submitting a query - for two different periods, the second period later than the first period. Like all statistical procedures one also has a sample of individuals, the sample being of sufficient size to provide statistically significant results. The rationale for the technique is that although the Web querying rate measurements of a group of users may vary enormously, for any particular user, the value of the later Web querying rate measurement will depend mostly on the value of the earlier Web querying rate measurement. The statistical procedure goes on to examine how justified one may be in using the earlier measurement to predict the later measurement. Clearly, if on average there were no change between the earlier and later periods, then the later Web querying rate measurement would equal the earlier Web querying rate measurement. Hence the extent to which this, later Web querying rate = earlier Web querying rate, is false for any sample of users indicates the existence of a phenomenon of change (in users' Web querying rates). Note that no presumption is made regarding the change effect being manifest as an increase (or decrease). Conditional regression is discussed by both Bijleveld et al. (1998) and Bock (1975).

For this investigation, the regression model $LATER_i = a\ EARLIER_i + e_i$ is used. Therefore the analysis is doing more than just examining whether the population of $LATER$ is different to the population of $EARLIER$. The conditional regression examines the changes in $LATER$, *given* a particular value of $EARLIER$. If $a = 1$, that is $LATER_i = EARLIER_i + e_i$, then the $LATER$ population differs from the $EARLIER$ population by only an error term. Hence there is no change. Conversely, when a is not (statistically) equal to 1 then change is present.

The conditional regression analysis is supported by an analysis of variance (ANOVA) for the Web host

conformance metric which examines variability in the data. The ANOVA model used considers three sources of variability, that among the users, that between the $EARLIER$ and $LATER$ data populations, and the interaction effect. Variability among the users and between the $EARLIER$ and $LATER$ populations would be expected. Variability in $EARLIER$ in particular would benefit computation of the conditional regression. Analysis of the interaction effect is of interest since it is interpreted as "user gaining experience".

Results

Descriptive summary

The investigation studied the Web information seeking activity of 206 users (98 male, 108 female). All the users were attending a UK college of Higher Education. Web information searching activity was monitored unobtrusively in order to compile anonymised client based Web transaction logs which were analyzed. The principal unit of analysis is the user information searching session which is taken to be all the URLs requested by the user within a single day of Web searching.

Both the number of Web searching sessions undertaken by each user, and the *duration* or calendar period of a user's Web searching varies. That is, some users are seen in the Web transaction log for only a short calendar period whereas other users appear over an extended calendar period. Figure 1 shows the histogram of the distribution of the number of sessions recorded by the users - the user's session count, while Figure 2 shows the duration of users' Web sessions, that is the inclusive period in days between a user's first and last session.

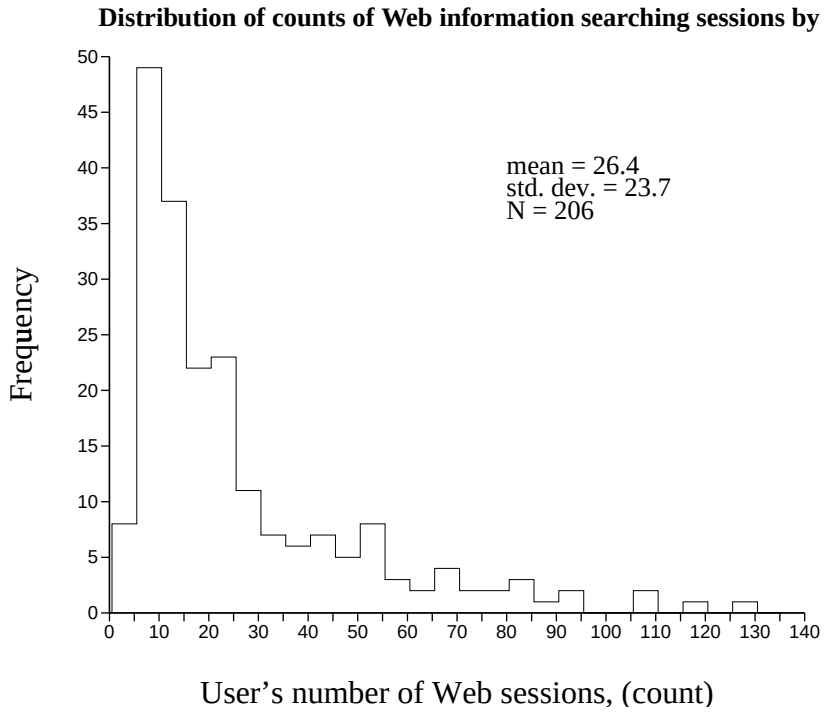

Figure 1: Histogram of user's session count

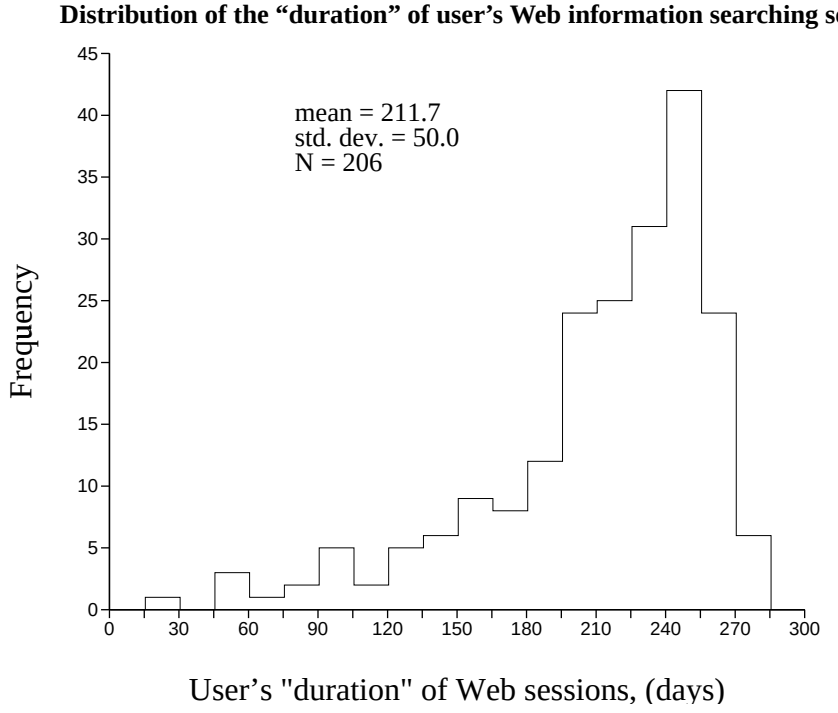

Figure 2: Histogram of user's "duration" (inclusive period between first and last) of sessions

The graph of the number of different Web hosts accessed by each user during the whole ten month survey period is shown in Figure 3.

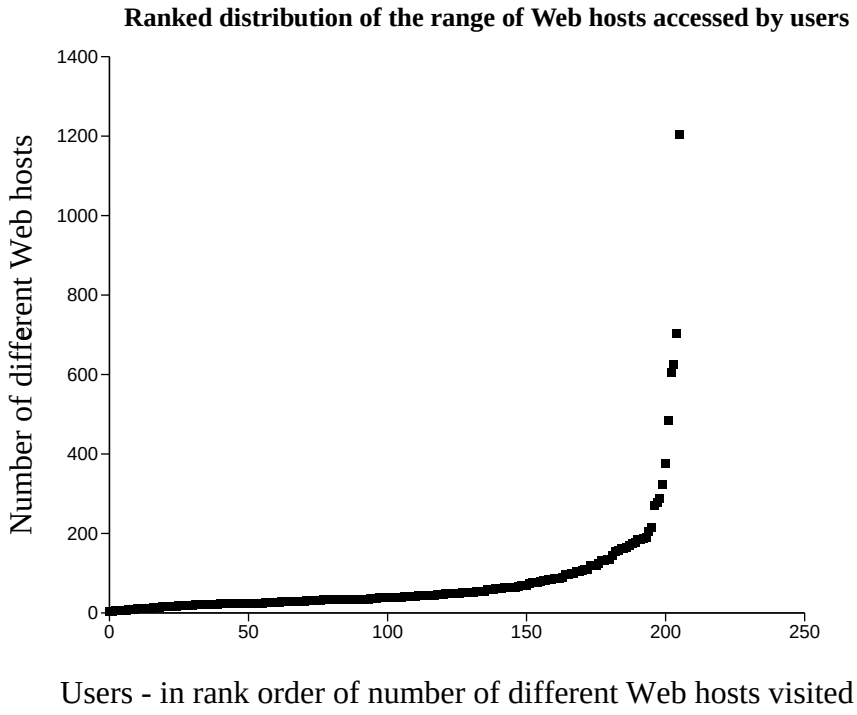

Figure 3: Graph of number of different Web hosts per user in rank order

These Figures show clearly that there is a broad range in these characteristics of users' Web information searching behavior. We next examine changes in user's Web searching as the user gains experience.

Longitudinal analyses

The longitudinal analyses uses a split-half technique. The half period split divides a user's Web information searching experience into two similar portions, earlier and later, in a way that corresponds to the real world of the users' Web information searching behavior rather than to any arbitrary calendar based method.

Change in the user's Web activity rate

For each user, Web *activity rate* = session count/duration, gives a measure of how concentrated in time are the user's sessions and combines the information shown in Figures 1 and 2 above. The activity rate

distinguishes between users who record the same number of sessions but over shorter or longer durations (and is equivalent to finding a user's reciprocal mean inter-session period or the average number of daily Web searching sessions per day).

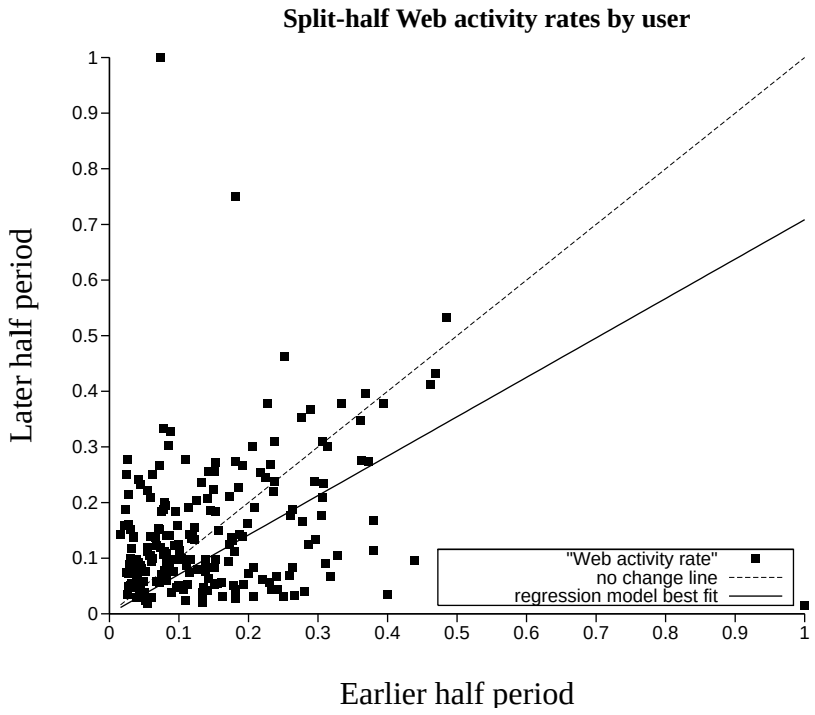

Figure 4: Graph of conditional regression of user's Web activity rates

Figure 4 shows a graphical plot of the users' Web activity rates analyzed according to the conditional regression model. The Figure also shows both the model's regression line and the *no change line*, that is the line predicting a user's later half period Web activity rate were there to be no change from the user's earlier half period Web activity rate. This line has a gradient of 1. The model estimate of the gradient of the regression line is between 0.6095 and 0.8075 (at 95% confidence level) which is therefore statistically significantly different from 1. Hence we may conclude that there is a change in users' Web activity rates between their earlier and later half periods.

Since, on average, a user's Web activity rate during their later half period is only 0.7 times the Web activity rate during their earlier half period, this implies that users access the Web less often, that is more sporadically, as they become more experienced.

Change in the user's querying rate

The conditional regression model, shown in Figure 5, investigates the change in a user's querying rate, that is the proportion of the user's Web information searching sessions when the user is searching more actively as opposed to using only a passive link-click based form of searching. As before, the no change and regression lines are also plotted.

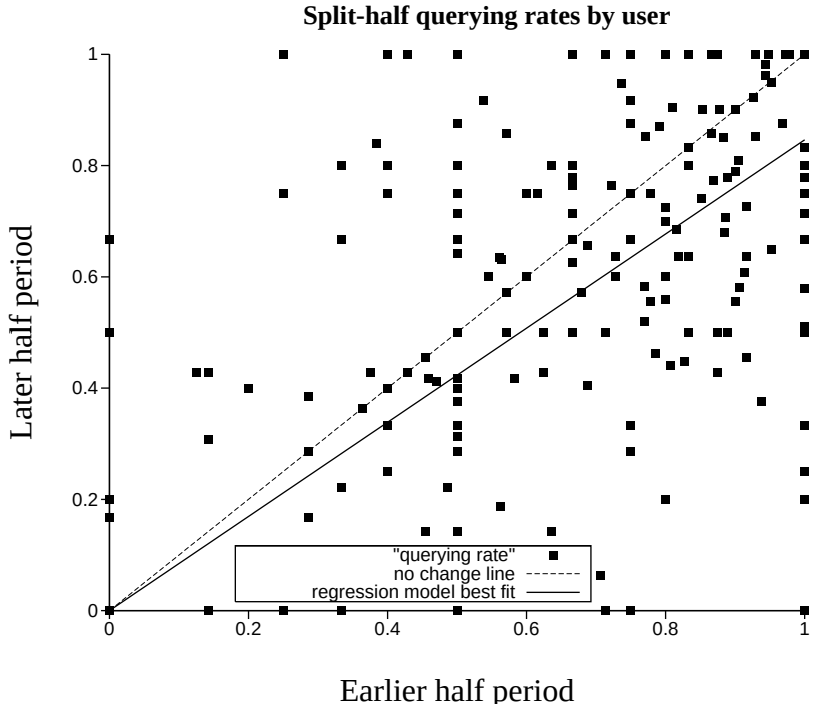

Figure 5: Graph of conditional regression of user's querying rates

The model estimate of the gradient of the regression line is between 0.7952 and 0.8976 (at 95 % confidence level). Hence, as before, there is a statistically significant change between the earlier and later half periods of users' Web querying rate when Web information searching.

Change in the user's Web host conformance

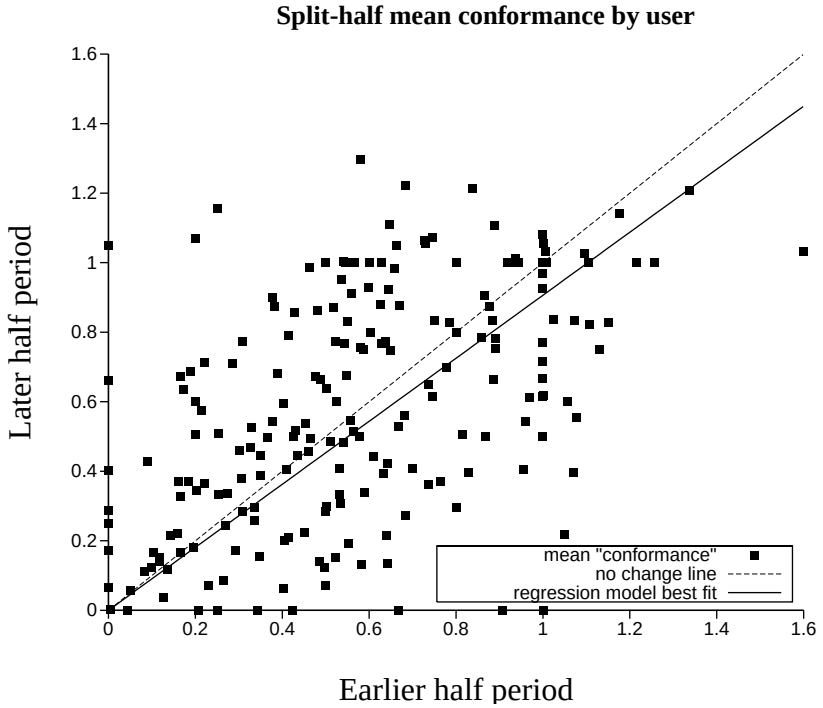

Figure 6: Graph of conditional regression of users' mean Web host conformance

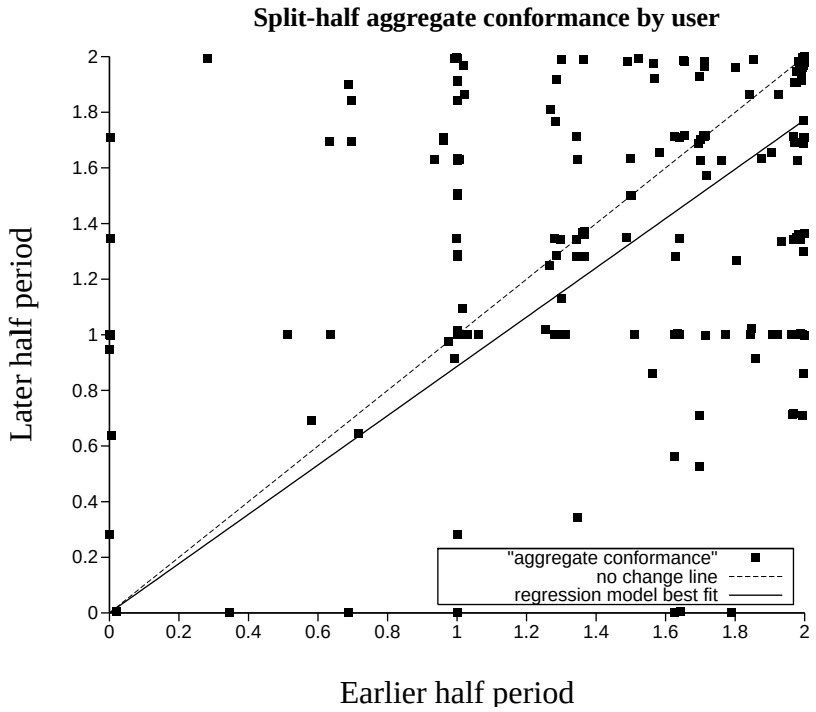

Figure 7: Graph of conditional regression of users' aggregate Web host conformance

The Web host conformance metric measures the degree to which a user chooses to access the more popular Web hosts, that is Web hosts accessed by most users. Figures 6 and 7 show the conditional regression model graphs for the users' mean Web host conformance and aggregate Web host conformance respectively.

Each user's mean conformance is found by taking the arithmetic average of the conformance measure for each of the user's sessions. Figure 6 is therefore showing for each user the conformance of the Web hosts accessed in a typical session during the user's earlier split-half period compared with a typical session during the user's later split-half period.

Figure 7 considers aggregate Web host conformance, that is, each user's choices of Web hosts during a half period are aggregated so that the conformance of a user omitting to access a popular Web host during one session will be compensated by that user including that Web host during another session within the same half period.

The apparent vertical and horizontal visual banding of data points about the axis values 1 in Figure 7 is a result of aggregating or clustering this version of the Web host conformance metric. Points close to 1 would indicate the user requesting just the most and a much lessor popular Web site (say $\theta = 1 + \frac{1}{128}$)

during the entire half-period. The graph shows that users are unlikely to do this and either request several of the 30 most popular sites or access the most popular site of the top 30 only during a half-period. (These latter users will also be searching other Web sites that because of their very low popularity are not included in the host conformance metric computation.)

The conditional regression model estimates (at 95% confidence level) for the gradients for the mean and aggregate Web host conformance measures are between 0.8405 and 0.9722, and between 0.8327 and 0.9400 respectively. In both cases the analysis therefore indicates that the users' level of Web host conformance changes between the two halves of their Web information seeking sessions (since the gradient value is statistically significantly different from 1, the gradient of the no change line).

Analysis of variance

The statistical analysis of variance (ANOVA) procedure can analyze the components of variability in the 5431 values of the Web host conformance metric. A standard two factor ANOVA model (using SPSS) where the factors are the user and the half period ("earlier" or "later") quantifies the variability due to the effects of each of the two factors alone and the factor interaction effect. The results are shown in the table.

Source of variance	F value	df	significance
User	11.5	205	$p < 0.01$
Half	7.0	1	$p < 0.01$
interaction	11.5	195	$p < 0.01$

Table 1: analysis of variance

There are statistically significant effects from both the user and half period factors. More importantly the ANOVA analysis identifies the interaction effect, which is experience, as being statistically significant. That is, the changes in the Web host conformance metric data due to experience are a real statistical phenomenon and are not due to chance random variation. 66% of the total variability in the Web host conformance metric is explained by the two factors and their interaction.

Discussion

The general notion of users becoming more systematic in their information seeking behavior as they become more experienced has appealed to investigators as they explain the findings of novice/expert comparisons. However definitions of systematic are elusive. Operationally, use has been made of, for example, the number of Boolean connectors, the length of search strings and the like. Previous studies have also imposed both an external definition of expert based on a user having greater than some threshold of prior experience, and a set of information tasks. Gross (1999) identifies the disadvantages of using imposed information tasks. This investigation uses both a real world definition of information tasks and a real world analysis of experience.

The results show change between the earlier and later periods of users' Web information searching behavior in that;

1. the activity rate (or mean inter-session period), that is how often the users used the Web, reduced,
2. users' querying rates reduced, that is users relied more on a passive link-clicking or browsing style of information searching, and
3. users' Web host conformance - both mean conformance and aggregate conformance decreased, that is users individually reduce their requests to the most popular Web hosts across the user sample.

These conclusions based on applying a conditional regression analysis are corroborated by the ANOVA analysis of the Web conformance measure for each daily user session of Web information searching.

As users become more experienced it is expected that their information searching behavior will change. However it is expected that users' behavior will become more systematic in the general sense of becoming more purposeful. Although these characterizations of user behavior are elusive, it was expected that a user's Web querying rate and Web host conformance metric would have increased (or possibly remained the same if the user has reached a plateau of expertise) which would be consistent with the user developing more active information searching behavior and requesting information from a common core (for the user sample) collection of Web hosts. This is not happening. Users adopted more passive or browsing styles of information searching and the range of Web hosts which they assess becomes less conformant or more eclectic as they became more experienced.

Possible confounding factors

Since the result found is the opposite of what was expected then some possible confounding factors which could produce the contra expected findings are discussed.

Information task

Many studies suggest that a user's choice of searching strategy is associated with the user's information task. Tasks which require simple fact finding promote more analytic style searching compared with more discursive information tasks. Therefore if there was a sufficient shift in the nature of the user's information tasks, that is a change in the information task environment, between users' earlier and later half periods then this might account for the unexpected observations. For example, since the user sample is a single cohort of students then the nature of the information tasks might be determined by particular tuition related factors. However these would be fixed in calendar time, that is the same tuition factors would apply to many students' Web information searching at the same point in time. However splitting user sessions into earlier and later periods is essentially a random procedure relating only to the sequence of a user's sessions and not to any external time reference. Both the number of user sessions and the duration of sessions (as shown in Figures 1 and 2) vary considerably so that Web searching during any particular calendar period will not occur consistently in either the earlier or later half periods.

Hence an underlying shift in the nature of the users' information tasks is not regarded as a feasible explanation for the contra expected findings.

Web evolution

The nature of the information system, in this case the Web, clearly influences the style of information seeking that users may display. It is accepted that the Web is subject to rapid growth and transformation, therefore could underlying changes in the Web confound the investigation's findings?

As with the information task, if the earlier and later periods were strongly associated with fixed points in time, then some change over time which changed the Web's information seeking affordances, for example by making full text journal article searching more accessible, could influence the results. However by a likewise argument this is not regarded as a possibility. In addition one would suspect that competition among Web hosts would cause the Web to evolve in a way that reduces rather than increases users' eclecticism.

Information "tourism"

Thomas (1998) in his longitudinal study of the variety of commands employed by computer users identifies that users go through an initial period during which they experiment and "try out" commands without absorbing them into their longer term repertoire. He follows Cooper (1991) in referring to this period as "tourism". The user sample for this investigation consists of students all in their second year of study who therefore all potentially had prior experience of Web information searching. Despite this one can hypothesize that the findings are distorted by "tourism", but if this were the case then Web conformance would be expected to increase not reduce.

Sample analysis bias

Could a small group of extreme users distort the findings? An unusual feature of the investigation is that the Web information searching behavior of (anonymised) individuals is monitored over time. This facilitates having the user as the prime unit of analysis rather than, for example, the session. Hence the findings about users are not distorted by some users recording very many more sessions than other users. Figure 3 shows that, in respect of the number of different Web hosts accessed, the number of extreme users is quite small and the analytic method precludes them having a disproportionate effect.

Conclusion

Theoretical implications

As yet, Web user information searching has received little investigative attention and analysis compared with other forms of electronic information systems. Therefore it may be that searching the Web information domain is sufficiently distinct that user characterizations of experience developed in non-Web information domains are not valid. Since the Web is becoming a ubiquitous medium for distributing information then we need to understand the phenomenon of Web information searching more fully in order to make information provision more effective. The study shows how standard tools of information retrieval and IR research can be used to undertake this task.

The study results are unexpected and may be particular to students although there are corroborating findings with other user groups. The results imply that Web users become more passive and more eclectic as they become more experienced. Not only do they fail to support the generally accepted view on the

effect of experience on user information searching behavior, they suggest that as Web users become more experienced they rely less on formal querying (typically using search engine) to obtain their Web based information. However as users become more experienced so their Web usage was more sporadic which suggests possible greater selectivity.

It appears that each user may inhabit an individual niche of Web hosts which becomes more distinctive as the user becomes more experienced. This conclusion is corroborated by Christ et al. (2001, p.2796) who found that " . . . in contrast to the exponential growth in Web sites available . . . there is actually a large decline in the average number of distinctive Web sites accessed", and Cockburn & McKenzie (2001) who report both an increase in revisitation rates, saying that " . . . even our subject with the lowest revisitation rate was revisiting pages more frequently than the overall mean of prior studies" (p.911), and individual distinctiveness " . . . there was a surprising lack of overlap in the pages visited by this fairly homogeneous community of users" (p.917).

This suggests that it is misleading to assume that there is a uniform large scale homogeneity in Web users' information searching behavior. The empirical evidence points towards individual users becoming increasingly distinctive as regards the Web sites which they use. A theoretical challenge is to identify characteristics that correlate with observed niche Web information searching behavior.

Implications for Web system design

Increases in Web traffic to major sites can be explained by the growth in the overall numbers of Web users rather than increased visits at the individual level (Christ et al., 2001). Hence future growth for major Web sites is limited as the flow of new users reduces and more experienced users adopt increasingly niche and passive Web information searching behavior. Web system designers may therefore wish to respond by developing a more selective dissemination of Web information to individual users which reflect their niche information attributes.

Acknowledgements

The author wishes to acknowledge the financial support of the UK Economic and Social Research Council, the technical co-operation and support of Cheltenham and Gloucester College of Higher Education which made possible carrying out this study, and the helpful comments of the anonymous reviewers.

References

Aalbersberg, I. J. (1994). A document retrieval model based on term frequency ranks. In Croft, W. B. & van Rijsbergen, C. J., eds., Proceedings of seventeenth annual international ACM SIGIR conference on research and development in information retrieval, Dublin, July 3–6, 1994, pp. 163–172, ACM SIGIR, London: Springer.

Abrahamson, A. D. (1998). Monitoring and evaluating use of the World Wide Web in an academic library: an exploratory study. In Preston, C. M., ed., Proceedings of the 61st ASIS annual meeting, Pittsburg, Pennsylvania, October 24–29, 1998, vol. 35, pp. 315–326, American Society for Information Science, Medford, New Jersey: Information Today for ASIS, [also online], available from: <url:http://webpages.csus.edu/~abramson/webstudy/abramson-webstudy.html>, [accessed 20 June 2001].

Bell, W. J. (1990). Searching behaviour: the behavioural ecology of finding resources. London: Chapman and Hall.

Berners-Lee, T., Masinter, L. & McCahill, M. (1994). RFC 1738: Uniform resource locators (URL). [Online], available from: <url:http://www.w3.org/Addressing/rfc1738.txt>, [accessed 24 October 2000].

Bijleveld, C. C. J., van der Kamp, L. J. T., Mooijaart, A., van der Kloot, W. A., van der Leeden, R. & van der Burg, E. (1998). Longitudinal data anlysis: designs models and methods. London: Sage.

Bock, R. D. (1975). Multivariate statistical methods in behavioral research. London: McGraw–Hill.

Borgman, C. L., Hirsh, S. G. & Hiller, J. (1996). Rethinking online monitoring methods for information retrieval systems: from search product to search process. Journal of the American Society for Information Science, 47(7), pp. 568–583.

Boyce, B. R., Meadow, C. T. & Kraft, D. H. (1994). Measurement in information science. London: Academic Press.

Carmel, E., Crawford, S. & Chen, H. (1992). Browsing in hypertext: a cognitive study. IEEE Transactions on systems, man and cybernetics, 22(5), pp. 865–884.

Catledge, L. D. & Pitkow, J. E. (1995). Characterizing browsing strategies in the World Wide Web. Computer networks and ISDN systems, 27(6), pp. 1065–1073.

Chen, H., Houston, A. L., Sewell, R. R. & Schatz, B. R. (1998). Internet browsing and searching: user evaluations of category map and concept space techniques. Journal of the American Society for Information Science, 49(7), pp. 582–603.

Choo, C. W., Detlor, B. & Turnbull, D. (1998). A behavioral model of information seeking on the Web: preliminary results of a study of how managers and IT specialists use the Web. In Preston, C. M., ed., Proceedings of the 61st ASIS annual meeting, Pittsburg, Pennsylvania, October 24–29, 1998, vol. 35, pp. 290–302, American Society for Information Science, Medford, New Jersey: Information Today for ASIS, [also online], available from: <url:http://choo.fis.utoronto.ca/fis/respub/asis98/>, [accessed 25 January 2001].

Christ, M., Krishnan, R., Nagin, D., Kraut, R. & Günther, O. (2001). Trajectories of individual WWW usage: implications for electronic commerce. In Proceedings of the 34th Hawaii international conference on system sciences, Maui, Hawaii, January 3–6, 2001, pp. 2794–2802, IEEE Computer Society, Los Alamitos: IEEE.

Cockburn, A. & McKenzie, B. (2001). What do Web users do: an empirical analysis of Web use. International journal of human–computer studies, 54(6), pp. 903–922.

Cooper, M. D. (1991). User skill acquisition in office information systems. Journal of the American Society for Information Science, 42(10), pp. 735–746.

Cooper, M. D. (2001). Usage patterns of a Web–based library catalog. Journal of the American Society for Information Science and Technology, 52(2), pp. 137–148.

Cove, J. F. & Walsh, B. C. (1988). Online text retrieval via browsing. Information processing and management, 24(1), pp. 31–37.

Cunha, C. R., Bestavros, A. & Crovella, M. E. (1995). Characteristics of WWW client-based traces. Technical report BU-CS-95-010, Computer Science Department, Boston University, [online], available from: <url:http://www.cs.bu.edn/techreports/95-010-www-client-traces.ps.Z>, [accessed 24 October 2000].

Fenichel, C. H. (1981). Online searching: measures that discriminate among users with different types of experiences. Journal of the American Society for Information Science, 32(1), pp. 23–32.

Fidel, R., Davies, R. K., Douglass, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., Miyagishima, B. K. & Toney, C. D. (1998). A visit to the information mall: Web searching behavior of high school students. Journal of the American Society for Information Science, 50(1), pp. 24–37.

Frants, V. I., Kamenoff, N. I. & Shapiro, J. (1993). One approach to classification of users and automatic clustering of documents. Information processing and management, 29(2), pp. 187–195.

Fu, Y., Sandhu, K. & Shih, M. (1999). Clustering of Web users based on access patterns. In Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, California, August 15–18, 1999: workshop on Web usage analysis and user profiling, August 15, 1999, ACM SIGKDD, New York: Association for Computing Machinery, [also online], available from: <url:http://www.acm.org/sigkdd/proceedings/webkdd99/papers/fu.htm>, [accessed 26 January 2000].

Goldstein, H. (1979). The design and analysis of longitudinal studies: their role in the measurement of change. London: Academic Press.

Graham-Cumming, J. (1997). Hits and miss-es: a year watching the Web. Computer networks and ISDN systems, 29(8–13), pp. 1357–1365.

Gross, M. (1999). Imposed queries in the school library media center: a descriptive study. Library and information science research, 21(4), pp. 501–521.

Harris, M. A. (1986). Sequence analysis of moves in online searching. Canadian journal of information science, 11(2), pp. 35–56.

He, D. & Göker, A. (2000). Detecting session boundaries from Web user logs. In Proceedings of the 22nd annual colloquium on information retrieval research, Cambridge, April 5–7, 2000, Information retrieval specialist group of the British Computer Society, [online], available from: <url:http://irsg.eu.org/irsg2000online/papers/source/he.pdf>, [accessed 31 January 2001].

Hölscher, C. & Strube, G. (2000). Web search behavior of Internet experts and newbies. Computer networks, 33(1–6), pp. 337–346.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge of the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44(3), pp. 161–174.

Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E. & Lukose, R. M. (1998, 3 April). Strong regularities in World Wide Web surfing. Science, 280(5360), pp. 95–97.

Iivonen, M. (1995). Searchers and searchers: differences between the most and least consistent searchers. In Fox, E. A., Ingwersen, P. & Fidel, R., eds., Proceedings of eighteenth annual international ACM SIGIR conference on research and development in information retrieval, Seattle, Washington July 9–13, 1995, pp. 149–157, ACM SIGIR, New York: Association for Computing Machinery.

Jansen, B. J., Spink, A. & Saracevic, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the Web. Information processing and management, 36(2), pp. 207–227.

Khan, K. & Locatis, C. (1998). Searching through cyberspace: the effects of link display and link density on information retrieval from hypertext on the World Wide Web. Journal of the American Society for Information Science, 49(2), pp. 176–182.

King, N. S. (1991). Search characteristics and the effects of experience on end users of PaperChase. College and research libraries, 52(4), pp. 360–374.

Kraut, R., Scherlis, W., Mukhopadhyay, T., Manning, J. & Keisler, S. (1996). The HomeNet field trial of residential Internet services. Communications of the ACM, 39(12), pp. 55–63.

Kurth, M. (1993). The limits and limitations of transaction log analysis. Library hi tech, 11(2), pp. 98–104.

Lazonder, A. W., Biemans, H. J. A. & Wopereis, I. G. J. H. (2000). Differences between novice and experienced users in searching information on the World Wide Web. Journal of the American Society for Information Science, 51(6), pp. 576–581.

Marchionini, G. (1995). Information seeking in electronic environments. Cambridge: Cambridge University Press.

Marchionini, G., Dwiggins, S., Katz, A. & Lin, X. (1993). Information seeking in full-text end-user-orientated search systems: the roles of domain and search expertise. Library and information science research, 15(1), pp. 35–69.

Mockapetris, P. (1987). RFC1034: Domain names: concepts and facilities. [Online], available from: `<url:http://www.ietf.org/rfc/rfc1034.txt>`, [accessed 28 November 2000].

Nesselroade, J. R. & Baltes, P. B., eds. (1979). Longitudinal research in the study of behavior and development. London: Academic.

Olson, M. A., Bostic, K. & Seltzer, M. (1999). Berkerley DB. In Proceedings of the USENIX annual technical conference, Monterey, California, June 6–11, 1999, USENIX Association, [online], available from: <url:http://www.usenix.org/events/usenix99/full_papers/olson/olson.pdf>, [accessed 21 June 2001].

Palmquist, R. A. & Kim, K. (2000). Cognitive style and on-line database search experience as predictors of Web search performance. Journal of the American Society for Information Science, 51(6), pp. 558–566.

Penniman, W. D. & Dominick, W. D. (1980). Monitoring and evaluation of on-line information system usage. Information processing and management, 16(1), pp. 17–35.

Peters, T. A. (1993). The history and development of transaction log analysis. Library hi tech, 11(2), pp. 41–66.

Pirolli, P. L. T. & Card, S. K. (1999). Information foraging. Psycological review, 106(4), pp. 643–675.

Pitkow, J. (1997). In search of reliable usage data on the WWW. Computer networks and ISDN systems, 29(8–13), pp. 1343–1355.

Pitkow, J. E. (1998). Summary of WWW characterizations. Computer networks and ISDN systems, 30(1–7), pp. 551–558.

Plewis, I. (1985). Analysing change: measurement and explanation using longitudinal data. Chichester: Wiley.

Qiu, L. (1993). Analytical searching vs. browsing in hypertext information retrieval systems. Canadian journal of information and library science, 18(4), pp. 1–13.

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing and management, 24(5), pp. 513–523.

Salton, G., Fox, E. A. & Wu, H. (1983). Extended Boolean information retrieval. Communications of the ACM, 26(12), pp. 1022–1036.

Salton, G. & McGill, M. J. (1983). Introduction to modern information retrieval. London: McGraw-Hill.

Schater, J., Chung, G. K. W. K. & Dorr, A. (1998). Children's internet searching on complex problems: performance and process analyses. Journal of the American Society for Information Science, 49(9), pp. 840–849.

Siegel, S. & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences. London: McGraw-Hill, second edition.

Spink, A., Wilson, T., Ellis, D. & Ford, N. (1998). Modeling users' successive searches in digital environments. D-lib magazine, 4(4), [online], available from: <url:http://www.dlib.org/dlib/april98/04spink.html>, [accessed 25 October 2000].

Spink, A. & Xu, J. L. (2000). Research note: selected results from a large study of Web searching. Information research, 6(1), [online], available from: <url:http://www.shef.ac.uk/~is/publications/infres/paper90.html>, [accessed 22 January 2001].

Sutcliffe, A. G., Ennis, M. & Watkinson, S. J. (2000). Empirical studies of end-user information searching. Journal of the American Society for Information Science, 51(13), pp. 1211–1231.

Tauscher, L. & Greenberg, S. (1997). How people revisit Web pages: empirical findings and implications for the design of history systems. International journal of human-computer studies, 47(1), pp. 97–137.

Thomas, R. C. (1998). Long term human-computer interaction: an exploratory perspective. London: Springer.

Yuan, W. (1997). End-user searching behavior in information retrieval: a longitudinal study. Journal of the American Society for Information Science, 48(3), pp. 218–234.