10-13-2009

Goal Attainment On Long Tail Web Sites: An Information Foraging Approach

James A. Mccart
University of South Florida

Goal Attainment On Long Tail Web Sites:

An Information Foraging Approach

by

James A. Mccart

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Information Systems and Decision Sciences
College of Business
University of South Florida

Co-Major Professor: Donald J. Berndt, Ph.D.
Co-Major Professor: Balaji Padmanabhan, Ph.D.
Joni L. Jones, Ph.D.
Richard P. Will, Ph.D.

Date of Approval:
October 13, 2009

Keywords: clickstream research, information foraging theory, web mining, information scent, data mining

Table of Contents

List of Tables

List of Figures

Goal Attainment on Long Tail Web Sites:
An Information Foraging Approach

James A. McCart

ABSTRACT

This dissertation sought to explain goal achievement at limited traffic "long tail" Web sites using Information Foraging Theory (IFT). The central thesis of IFT is that individuals are driven by a metaphorical sense of smell that guides them through patches of information in their environment. An information patch is an area of the search environment with similar information. Information scent is the driving force behind why a person makes a navigational selection amongst a group of competing options. As foragers are assumed to be rational, scent is a mechanism by which to reduce search costs by increasing the accuracy on which option leads to the information of value.

IFT was originally developed to be used in a "production rule" environment, where a user would perform an action when the conditions of a rule were met. However, the use of IFT in clickstream research required conceptualizing the ideas of information scent and patches in a non-production rule environment. To meet such an end this dissertation asked three research questions regarding (1) how to learn information patches, (2) how to learn trails of scent, and finally (3) how to combine both concepts to create a Clickstream Model of Information Foraging (CMIF).

The learning of patches and trails were accomplished by using contrast sets, which distinguished between individuals who achieved a goal or not. A user- and site-centric version of the CMIF, which extended and operationalized IFT, presented and evaluated hypotheses. The user-centric version had four hypotheses and examined product purchasing behavior from panel data, whereas the site-centric version had nine hypotheses and predicted contact form submission using data from a Web hosting company.

In general, the results show that patches and trails exist on several Web sites, and the majority of hypotheses were supported in each version of the CMIF. This dissertation contributed to the literature by providing a theoretically-grounded model which tested and extended IFT; introducing a methodology for learning patches and trails; detailing a methodology for preprocessing clickstream data for long tail Web sites; and focusing on traditionally under-studied long tail Web sites.

Chapter 1

Introduction

Understanding the browsing behavior of users at Web sites has been the objective of much of the research employing data about users' Web usage (commonly known as "clickstream data"). Especially salient has been the investigation of factors relating to choice behavior, where choice is typically concerned with the purchase of a product (Bucklin et al., 2002). Besides having a general understanding of why users behave the way they do, such knowledge also forms the basis for developing mechanisms to influence choice. For example, to steer a visitor towards a purchase, dynamic on-the-fly changes may be made to a Web site in terms of its "...pages, link choices, promotional interventions, and prices and product assortments" (Bucklin et al., 2002, pg. 252).

Such a general understanding of factors affecting choice; however, has been difficult to obtain. In part, the difficulty arises because conceptual research focusing on the theories and ideas which provide an explanation of a user's behavior has been limited (Bucklin et al., 2002). This lack of a theoretical base negatively impacts the ability of the results from clickstream research to be reconciled, synthesized, and thus provide a clearer picture of those factors.

Finding an appropriate theory to use is challenging in light of the type of data available. Clickstream data provides information on the actions of a user (e.g., what pages were visited, how much time was spent at a site), but nothing else. A person's attitudes, emotions, intentions, and other such concepts are unknown. However, many theories examining an individual's behavior in information systems research rely on such concepts as attitudes and intentions (e.g., Theory of Planned Behavior (Ajzen, 1991)) and thus are not appropriate to use. Therefore, a theory is needed which can (1) explain behavior based on a user's action and (2) be appropriately applied to the clickstream domain.

Within the last decade, a theory called Information Foraging Theory (IFT) has emerged which explains the searching behavior of individuals as they hunt for information (Pirolli and Card, 1999). The thesis of IFT is that an individual is driven by a metaphorical sense of smell that guides them

through patches of information in their environment based on their information goal (i.e., what they are trying to accomplish) (Pirolli, 2007). As they "forage", individuals evaluate whether to continue browsing in their current patch of information or leave to hunt for another one. Central to this theory are the concepts of information patches and information scent. Information patches are distinct areas of the search environment which differ in informational content. Information scent is the driving force of why a person makes a navigational selection amongst a group of competing options.

IFT itself builds on more established theories such as Optimal Foraging Theory (OFT) (Stephens and Krebs, 1986) and the Adaptive Control of Thought-Rational Theory (ACT-R) (Anderson et al., 2004). OFT is an ecological theory concerned with explaining the foraging behavior of animals as they hunt for food. OFT assumes each animal goes through a search–encounter–decision process as they forage, with the goal being to maximize net energy gained. To maximize energy, the animal is faced with the decision of which prey to eat or how long to forage in a patch. OFT is used to explain the behavioral elements of people foraging for information.

ACT-R is a psychological theory of the human mind that includes the cognitive architecture and process by which cognition works. IFT uses a production rule system from ACT-R to determine probabilistically which action is selected based on its utility within the context of a user's current goal. For example, an action to click on a hyperlink may be chosen over backing up to a previously visited page because following the hyperlink may be more likely to lead to the information being sought. ACT-R is used to explain at a cognitive level why actions are performed.

IFT was originally developed to be used in a "production rule" environment, where a user would perform an action when the conditions of a rule were met. However, the use of IFT in clickstream research requires conceptualizing the ideas of IFT in a non-production rule environment. In essence, this requires utilizing user action to infer the cognitive process and thus the reasoning behind the observed behavior. To meet such an end this dissertation describes how information patches and trails of information scent can be learned from clickstream data. However, the main focus of this dissertation is to determine how the concepts of IFT can be used to build a clickstream model of information foraging (CMIF). The model relies on measures derived from clickstream data representing IFT concepts to explain goal achievement at "long tail" Web sites that have limited traffic. Goal achievement is from the perspective of the online firm and consists of something the firm

would like to happen at their Web site (i.e., a choice). This dissertation examines Web sites where the goal is the purchase of a product or the submission of a contact form.

The term “long tail” refers to a Web site that resides in the tail of a power law distribution (Anderson, 2006). Figure 1 shows a hypothetical power law distribution illustrating Web sites and their popularity in terms of the number of visits they received[1]. The head of the curve (darkly shaded portion) represents the most popular Web sites such as Amazon.com and eBay.com. The long drawn-out tail of the curve (lightly shaded portion) extends to include all other Web sites.

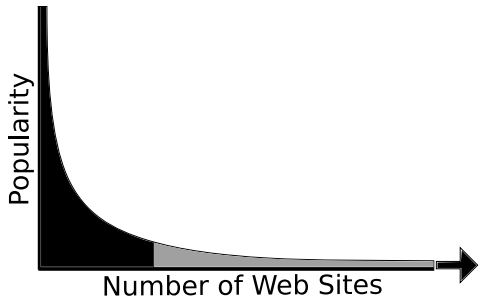


Figure 1.: Power Law Distribution

The decision to analyze a user’s behavior at long tail Web sites was motivated by the ability of IFT to guide analysis. Compared to sites in the head, long tail Web sites have significantly smaller amounts of data, which is precisely where theory can help guide analysis the most. Lacking theory, analysis would require large amounts of data to work well with techniques commonly used such as data mining. Long tail Web sites by their very nature are prohibitively sparse in data which hamper the application of such an exploratory approach.

The remainder of this chapter is organized as follows. First, the research questions guiding this dissertation are introduced in §1.1. A brief discussion of the contributions of this dissertation are given in §1.2. Finally, §1.3 provides a brief overview of the structure of this dissertation.

[1]The power law distribution of Web sites and traffic has been previously confirmed through empirical study (Adamic and Huberman, 2001) and simulation (Kavassalis et al., 2004). Power-law distributions have also been observed in numerous other instances such as the sales of products (Anderson, 2006); frequency of word usage in English text; number of telephone calls received; frequency of family names in the United States; and citations of academic papers (Newman, 2005).

1.1 Research Questions

The following subsections describe the research questions guiding this dissertation. The first research question is in relation to the concept of information patches. The second research question more fully explores the concept of information scent. Finally, the third research question brings all the concepts of IFT together to develop a clickstream model of information foraging.

1.1.1 Research Question 1 – Learning Patches

An information patch is defined as an area of the search environment with similar information (Pirolli, 2007). Within a Web-context, what constitutes a patch is dependent on the level of analysis being examined. At a high-level of analysis, an entire Web site can be considered a patch. When examined from a finer-grained level of analysis, each individual page of a Web site can also be considered a patch. While such conceptualizations of a patch are straightforward, they are effectively being defined by the creator of the content rather than the user.

The Web, however, is a pliable environment where foragers have the choice of what material to view. Effectively, this allows a forager to define their own information patch that is uniquely relevant to their goal. Such patches may consist of a group of Web pages, which individually may mean very little, but when combined provide an area of the search environment that is seen as valuable to the user. Therefore, the first research question attempted to discover how such patches can be learned.

Research Question 1: *How can information patches be learned from a long tail Web site?*

Although each user is free to define patches of value as they see fit, certain patterns of patches may emerge among foragers with similar information goals. From the viewpoint of the online firm, knowing who values what patch can provide insights into the information goal of the forager. By categorizing patches as valuable to goal-achievers or non-goal-achievers, the firm may be able to better explain goal achievement at long tail sites dependent on what patches were visited by a user. Therefore, a measure was also developed which quantified a user's visitation of valuable *goal patches*.

1.1.2 Research Question 2 – Learning Scent Trails

Information scent is the driving force behind why a person makes a navigational selection amongst a group of competing options. As foragers are assumed to be rational, scent is a mechanism by which foragers' reduce their search costs by increasing their accuracy on which option leads to the information of value (Pirolli, 2007). Based on the information goal of a forager, each hyperlink on a Web page gives off a scent. The higher the scent the more likely the page that is being linked to may contain the information being sought. Similar to a bloodhound that follows a scent trail over distances to find an item of interest, a forager also follows a scent trail to find the information they seek over multiple Web pages. The second research question sought to explain how scent trails may be learned.

Research Question 2: *How can information scent trails be learned from a long tail Web site?*

Similar to the learning of patches, each user may have their own scent trail. However, patterns may exist from fragments of scent trails that emerged among foragers with similar information goals. These fragments of scent trails are of value to the online firm in distinguishing between possible goal-achievers and non-goal-achievers. When a user follows these known fragments of scent trails it may provide clues into their information goal and thus help in explaining goal achievement at long tail sites. Thus, a measure was developed which computed the following of *goal scent trails*.

1.1.3 Research Question 3 – Clickstream Model of Information Foraging

The previous two research questions examined the concepts of information scent and patches individually. However, the real value of IFT is its ability to combine aspects of a user's search environment (i.e., patches) and their actions (i.e., scent) together. Thus the main focus of this dissertation and the final research question was how these concepts could be combined using clickstream data to infer goal achievement.

Research Question 3: *How can information foraging theory and clickstream data be used to explain the achievement of a goal at a long tail Web site?*

To answer the third research question, two versions of a clickstream model of information for-

aging (CMIF) were created which used clickstream metrics to represent the concepts of information scent and patches. The user-centric (UC) model exploited user-centric data (Padmanabhan et al., 2001) about a forager's entire browsing behavior to explain goal achievement at a long tail Web site. This model compared a forager's behavior across multiple Web sites. However, due to user-centric data typically being aggregated at the session level, the model lacked depth at individual Web sites.

Since data about a user's entire clickstream over multiple sites is rarely available to an online firm, a site-centric (SC) version of the model employing site-centric data (Padmanabhan et al., 2001) was also developed. Page-level data made the site-centric model capable of analyzing patches at all levels of analysis along with information scent at a Web site.

1.2 Contributions

Listed below are the major contributions of this dissertation.

First, this dissertation demonstrated how IFT could be used as a theoretical basis for clickstream research. Through the creation of two versions of a clickstream model of information foraging, the concepts of IFT were quantified outside of a production rule environment. In addition, the CMIF not only operationalized the core concepts of IFT, but also extended the theory by introducing memory, forager-independent valuation of patches and trails, along with refined definitions of scent (e.g., strict and relaxed scent). Once tested, many of the core concepts of IFT were supported, as were many of the theoretical extensions. Thus, this dissertation not only demonstrated the ability of IFT to explain goal achievement, but it also introduced theoretical extensions which provided a more in-depth explanation of goal behavior.

This dissertation also presented a methodology on how to learn patches and scent trails using not only significant, but also supported contrast sets. Measures were also created which quantified a forager's visitation of patches and following of trails. The metrics measured the most valuable, last, and summation of all patches and trails that were visited or followed. For those Web sites within the CMIF that discovered patches and trails, the measures were capable of distinguishing goal from non-goal sessions according to a forager's visitation and following behavior.

The third contribution was a methodology that detailed how to preprocess datasets with long tail Web sites. In particular, a separate user- and site-centric methodology was presented which high-

lighted the unique challenges associated with preprocessing each dataset. For example, a process was provided for the site-centric dataset about how to locate and select a single definable goal on Web sites which have more than one available goal.

Finally, due to the presence of IFT guiding analysis, traditionally under-studied long tail Web sites were able to be examined even in light of their sparse datasets. As far as can be determined, this dissertation is the first to empirically study goal achievement on long tail Web sites.

1.3 Dissertation Structure

The structure of the remaining chapters of this dissertation is outlined below. Each item in the list provides a brief summary of the main purpose of each chapter.

Chapter 2. Literature Review – An overview of prior clickstream research along with the datasets and metrics used in that research.

Chapter 3. Theory – A detailed explanation of information foraging theory along with the two theories IFT draws from: ACT-R and OFT.

Chapter 4. Hypotheses – The hypotheses for both the user- and site-centric versions of the CMIF (third research question). In addition, an explanation of the extensions to IFT is provided.

Chapter 5. Methodology – A separate methodology for the user- and site-centric versions of the CMIF is presented that covers the data used, how measures were calculated, and finally how the hypotheses were tested. The appendix contains a description of how to learn patches and trails (first two research questions).

Chapter 6. Datasets – A detailed explanation of the series of preprocessing steps each dataset went through to obtain a final dataset.

Chapter 7. Results – A listing and discussion of the results for each of the three research questions. Descriptive statistics are provided about learned patches and trails. In addition, statistical tests and a discussion of each of the hypotheses for the third research question are also provided.

Chapter 8. Temporal – An alternate time-sensitive representation of the site-centric CMIF. The methodology, results, and discussion are provided for the seven tested hypotheses.

Chapter 9. Conclusion – Summarizes this dissertation and provides limitations, contributions, and directions for future research.

Chapter 2

Literature Review

This chapter provides a summary of prior research which has focused on the behavior of visitors at Web sites using clickstream data. A brief list of terms commonly used throughout this dissertation are provided first in §2.1. Then prior research is summarized and classified by research focus in §2.2. Table 1 lists the general research questions and results of prior research, while §2.2.1–2.2.4 gives more in-depth descriptions of the literature. The datasets and metrics used in each study are then discussed in §2.3 and §2.4, respectively.

2.1 Terminology

In order to be clear and consistent, definitions of terms commonly used throughout this dissertation are provided below. Each bolded term is followed by its definition. If any synonymous terms exist they are italicized in parentheses immediately following the bolded term.

Path – sequence of Web pages viewed during a session or user session.

Sector – a collection of Web sites with products, services, and/or information which are of a similar nature (e.g., food).

Session – a time-contiguous sequence of Web page views at the same Web site for the same visitor.

User Session – a time-contiguous sequence of Web page views at any number of Web sites for the same visitor.

Visitor (*user*) – a person making Hypertext Transfer Protocol (HTTP) requests at a single Web site or multiple Web sites.

Web page (*page*) – a file written in Hypertext Markup Language (HTML) containing information that is viewable via a Web browser (e.g., index.html).

Web site (*site*) – a collection of Web pages housed under the same top- and second-level domain name (e.g., amazon.com).

2.2 Prior Research

In keeping with prior frameworks, the objective of a visitor at a site can be classified as browsing or purchasing (Bucklin et al., 2002). A browsing objective reflects how a visitor may navigate within a site (Bucklin and Sismeiro, 2003), across multiple sites (Park and Fader, 2004), or how site visits evolve over time (Moe and Fader, 2004a). Conversely, a purchasing objective is interested in discovering factors which affect a visitor's propensity to purchase (Sismeiro and Bucklin, 2004). However, the purchasing objective can be seen as a specific instance of the more general goal achievement objective as many sites have purposes other than purchasing (e.g., filling in a contact form, posting a message, responding to a survey). Therefore, the objective of a visitor can be classified as browsing, purchasing, achieving a goal[1], or exploring multiple objectives simultaneously (Moe, 2003).

Table 1 categorizes past studies by which objective the research was examining and then summarizes the research questions and results obtained. A more thorough description of prior research is provided in the subsections following the table.

[1] Although purchasing is a subset of goal achievement it is retained as a separate objective since numerous studies specifically examine purchasing behavior.

Table 1: Prior Literature: Results

Article	Research Question / Purpose	Results
MULTIPLE OBJECTIVES		
Kalczynski et al. (2006)	How well do clickstream-complexity measures predict task completion?	Two Web site-independent clickstream-complexity measures representing the linearity and density of a session were found to perform the best with accuracies between 65% and 93% depending on the task and site.
Moe (2003)	What visitor behavior can be uncovered from the pattern and type of pages viewed?	Four groups of visitors differing in search behavior and purchasing horizon were found, along with a fifth group of non-serious visitors. The purchase probability of each group differed depending on how immediate the purchase was and how directed the browsing behavior was.
BROWSING		
Bucklin and Sismeiro (2003)	Do visitors change the way they browse a Web site at the session or site level?	Visitors did dynamically change their browsing behavior at both the session and site level. Within a session browsers exhibited lock-in as they browsed deeper into a Web site. Across sessions a learning effect was observed which reduced the number of pages viewed, but not the duration spent on each page.

Continued on Next Page...

Table 1: Prior Literature Results – Continued

Article	Research Question / Purpose	Results
Danaher et al. (2006)	What factors affect visit duration?	Age interacted with gender, Web site functionality, and the graphical content of a site negatively with regards to the duration spent on a Web site. Age interacted positively increasing duration for older visitors for higher levels of text and advertisements on a Web site.
Johnson et al. (2004)	Does reduced search cost lead to increased search?	Overall search levels were low across the three sectors examined. Browsing behavior was also found to differ depending on sector and level of activity.
Moe and Fader (2004a)	To model individual-level evolving visit patterns over time.	Examining data at an individual-level contradicted aggregated visit patterns. More frequent visits and an increase in visiting rates increased visitors' probability of purchasing.
Park and Fader (2004)	Understand cross-site visiting behavior at the individual level.	An ability to predict when a visitor will first visit a Web site given their visiting pattern at another site.
Zhang et al. (2006)	How does search cost, product characteristics, previous search behavior, and consumer characteristics affect search depth?	Lower search costs and prior search behavior were positively correlated with search depth. Price and consumer characteristics were positively correlated to search depth for only certain product types.

Continued on Next Page...

Table 1: Prior Literature Results – Continued

Article	Research Question / Purpose	Results
PURCHASING		
Moe and Fader (2004b)	To model individual-level dynamic conversion behavior.	The individual-level model contradicted aggregated conversion trends. Over time the overall purchase probability of a visitor decreased, repeat visits had less of an impact on purchasing, and visitor experience raised the purchasing threshold.
Montgomery et al. (2004)	Can the path a visitor takes through a Web site help predict purchase?	Future paths were predicted with greater accuracy by the model using paths and by allowing search behavior (i.e., exploratory, directed) to change during a session. Purchase prediction was 10% and 21% accurate after a visitor viewed one page and six pages, respectively.
Padmanabhan et al. (2001)	What are the implications of using site-centric (i.e., incomplete) data versus user-centric (i.e., complete) data?	Models using user-centric data outperformed models using site-centric data by a wide margin. Using site-centric data can lead to erroneous results since significant metrics in site-centric models may no longer be significant in user-centric models.
Sismeiro and Bucklin (2004)	Does viewing the purchasing process as a series of tasks increase prediction accuracy?	The multi-task model outperformed the competing single task models supporting the series of tasks concept. The model metrics differed in effect sign, size, and significance between tasks indicating some metrics were better predictors of some tasks over others.
Van den Poel and Buckinx (2005)	How well do different types of metrics predict purchases?	Detailed clickstream metrics, which were divided according to the underlying content of the page (e.g., product information, community pages), were found to be the most important predictors of purchase.

Continued on Next Page…

Table 1: Prior Literature Results – Continued

Article	Research Question / Purpose	Results
GOAL ACHIEVEMENT		
Chatterjee et al. (2003)	To model a visitor's probability of clicking a banner advertisement.	Advertisements exhibited "wearout" such that multiple exposures reduced the probability of a visitor clicking an advertisement. Infrequent visitors were also more likely to click on a banner advertisement than frequent visitors.

2.2.1 Multiple Objectives

Moe (2003) created a two-dimensional typology and sought to discover metrics which helped categorize the within-session shopping strategies of visitors. The first dimension of the typology, search behavior, was dichotomized into following a directed or exploratory pattern[2] (Janiszewski, 1998). Directed searching occurs when a visitor has a particular goal or product in mind (Rowley, 2000). Exploratory search, on the other hand, takes an undirected approach where the visitor may not be attempting to locate a particular product or meet a specific goal. The time horizon, in which the expected purchase is to take place, either immediately or in the future, was the second dimension of the typology.

As seen in figure 2, four categories of shopping strategies emerged from the typology: directed buying; search and deliberation; hedonic browsing (i.e., exploratory or stimulus-driven browsing); and knowledge building. Each strategy was expected to have a unique pattern of the type, variety, and number of repeat viewings of particular types of pages.

Purchasing Horizon	Search Behavior	
	Directed	Exploratory
Immediate	Directed buying (20.0%)	Hedonic browsing (1.4%)
Future	Search/deliberation (6.6%)	Knowledge building ($< 0.1\%$)

Figure 2.: Shopping Strategy Typology (Moe, 2003)

Using seven weeks of data from a nutritional supplement store, Moe (2003) empirically tested the typology using cluster analysis and found all four theorized categories were present along with a fifth category of non-serious visitors[3]. The two most important metrics found for discriminating

[2] Bloch et al. (1986) created a framework for consumer information search and also delineated between two search behaviors, pre-purchase and ongoing search. Pre-purchase search, which was defined as seeking to facilitate decision making about a particular goal, maps to the directed search behavior. Ongoing search maps to the exploratory search behavior and was defined as searching that is independent of a goal.

[3] Visitors in the fifth category, on general, viewed two pages and spent a short amount of time on each page. Due to the limited browsing behavior exhibited on the site by these visitors before leaving they were not considered as having a serious interest in the site. Nicholas et al. (2007) termed those non-serious visitors as 'bouncers' who go from site to site without deeply penetrating or frequently returning to the Web site of interest.

between shopping strategies within a session was the number of different category pages viewed and the maximum number of times a product page was viewed (Moe, 2003). Figure 2 also contains the conversion rate of each category, in parentheses, which was found to range from $< 0.1\%$ to 20.0%.

Examining the behavior of visitors performing purchasing and information-seeking tasks over five Web sites, Kalczynski et al. (2006) used the navigational complexity of a visitor's session to help predict the completion of tasks. The central idea of navigational complexity is the correspondence with an underlying search behavior (e.g., Moe, 2003) where, for example, a less complex session is associated with a directed search behavior whereas greater complexity in a session points toward an exploratory search behavior. Using graph theory, each visitor's session was decomposed into a clickstream graph which represented the Web pages and links traversed within a Web site and allowed for the calculation of navigational complexity.

A total of 485 sessions, in a controlled experiment, attempted to complete three purchasing and three information-seeking tasks with the overall success rates for the tasks varying from 8.8% to 56% (Kalczynski et al., 2006). Two clickstream graph-complexity metrics representing the linearity and density of a session were used in binary logistic regression models for each task. Overall, the models correctly classified a session between 64.9% and 93.1% of the time depending on the Web site and task[4] (Kalczynski et al., 2006).

2.2.2 Browsing

Moe and Fader (2004a) explored the pattern and evolution over time of a visitor's browsing behavior at the individual-level. The authors argued that aggregating browsing behavior at the site-level to create general traffic patterns may lead to a false understanding of the complete browsing behavior occurring at a site (Moe and Fader, 2004a). For instance, aggregated data may indicate an upward trend in both the number of visitors and rates of visits to a site. The inclusion of new visitors may however, be masking a decline in visiting rates for experienced visitors (i.e., established customers).

Moe and Fader (2004a) used eight months of user-centric data focusing on Amazon.com and

[4]The model with 93.1% accuracy was for the task with only 8.8% success. As only the overall accuracy and not the specificity and sensitivity were provided the practical benefit of the model is unknown, Kalczynski et al. (2006) acknowledged this limitation.

CDNow.com to validate a nonstationary evolving visit model. The model took into account an individual's heterogeneity, visiting rate, and evolution of visiting rates over time. Compared against an exponential-gamma timing process, which did not allow for change over time, the evolving visit model was more accurate in estimating the likelihood and when a visitor would return to a site (5% overprediction versus 37%) (Moe and Fader, 2004a). In addition, the distribution of visiting rates did show a decline in visiting rates for experienced visitors which contradicted the aggregated trends. Furthermore, more frequent visits and an increase in visiting rates were found to be significant in terms of a visitor's probability of purchasing (16.6% vs. 11.1% and 5.5% versus 2.4%, respectively) (Moe and Fader, 2004a).

Also concerned with aggregated statistics being used to infer visitors' browsing behavior, Bucklin and Sismeiro (2003) created an individual-level model of browsing behavior within a Web site. The first aspect of the model accounted for a visitor's decision to continue browsing the site or exit the site. The second aspect was concerned with the duration of time a visitor spent on each individual page. Using a type II tobit model and one month of data from a commercial automotive Web site, four distinct browsing behaviors were identified: a learning effect, within-site lock-in, time-constraints, and a cost-benefit view.

The results of the first behavior, learning effect, was consistent with prior research showing the overall duration of sessions decreased with each subsequent session (Johnson et al., 2003). Although the overall session duration and number of pages viewed decreased, the duration spent on each page did not significantly differ from previous sessions (Bucklin and Sismeiro, 2003). The second behavior was based on the concept of lock-in (Johnson et al., 2003; Zauberman, 2003); however, in this context the lock-in corresponded to a visitor becoming more engrossed as they continued to browse a Web site within the same session instead of over time. The results supported this idea of within-site lock-in since the amount of time spent viewing each page increased as the number of pages viewed in a session increased (Bucklin and Sismeiro, 2003).

Time-constraints, the third behavior, showed the probability of a visitor staying on the Web site decreased as the overall session duration increased. The final behavior demonstrated visitors' likely performed some type of cost-benefit analysis since a page with greater amounts of information increased a visitor's probability of staying on the Web site. However, the probability of a user leaving the site can also increase with greater levels of information. For example, reading all

the information on a page may result in longer page durations which translate into longer session durations. Due to time-constraints, longer session durations then leads to a greater probability of a user leaving the Web site (Bucklin and Sismeiro, 2003).

Echoing the concerns of Padmanabhan et al. (2001) about using incomplete data, Park and Fader (2004) posited the timing and frequency of future visits to a site can be better explained by examining visiting behavior at other sites. Specifically, the browsing behavior in terms of visit timing and visit rates compared to other sites can be examined. For instance, one visitor may have high visit timing in which a visit to one site is followed shortly by a visit to another site. A different visitor may have a high visit rate where the number of visits to each site is similar, regardless of the coincidence of visit timing. The relationship of both these concepts to a visitor's browsing behavior can be used to predict future visits to a site of interest.

Using a multi-variate timing mixture model with closed-form analytic expressions, Park and Fader (2004) looked at the browsing behavior of visitors from two pairs of sites within the book and music sectors. Four models were compared, which differed based on if correlation in visit timing and rates were accounted for, with the proposed model accounting for both correlations. The proposed model was found to provide the best fit and performed well when long spaces of time occurred between visits (Park and Fader, 2004). However, when visits to different sites occurred on the same day, the proposed model failed to perform as well. The proposed model also outperformed the other three models in identifying zero class customers (i.e., customers who have not visited the Web site) who become non-zero class customers (i.e., customers who will visit the Web site) in the future (Park and Fader, 2004).

Since the average duration a visitor spends on a Web site is a component of the stickiness of that site (i.e., ability to attract and keep the interest of a visitor) (Bhat et al., 2002), Danaher et al. (2006) set out to uncover the factors that affect visit duration. The resulting model took into account two sources of individual-level heterogeneity in the form of demographics and a visitor's situational characteristics for a particular visit to a site (e.g., weekday versus weekend visit, number of previous visits). Site-level heterogeneity included measures of the textual, graphical, and advertising content of a Web site. Measures representing the background complexity and overall Web site functionality[5] were also included in the model.

[5]Functionality was measured as the average of 19 binary items indicating the presence or absence of features on the Web site such as "…online help, search functions, site maps, user registration, e-mail contact availability, chat rooms,

Using a month of panel data from 1,655 panelists for the 50 most-visited Web sites in the dataset, the developed model demonstrated that all three sources of heterogeneity were significant in explaining duration. Although all significant, visitors' situational characteristics accounted for almost 80% of the variance explained (Danaher et al., 2006) providing support that clickstream metrics in the absence of demographics and Web site characteristics can explain a substantially part of visitors' behavior. In terms of specific metrics, age was found to interact significantly with almost all of the demographic and Web site-specific metrics. For instance, the functionality of a Web site and age interacted such that an increase in functionality decreased the duration a visitor spent on the site the older the visitor was. The opposite relationship was found between age and advertising content where visit duration increased for older visitors when visiting Web sites with more advertisements.

Due to the relatively costless nature of searching on the Internet, Johnson et al. (2004) sought to answer the question does reduced search costs lead to increases in search behavior? To answer that question search behavior was operationalized into three components consisting of the depth, dynamics, and activity of search. The resulting Hierarchical Bayesian model accounted for a household's visitation of multiple sites within the same sector (depth), change in search behavior over time (dynamics), and amount of overall activity in a sector (activity).

Focusing on three sectors (books, music, and air travel) the search behavior of households at 51 of the most visited Web sites (13 books, 16 music, and 22 air travel) within the dataset were analyzed. Consistent with prior research, it was found that overall households searched very little (Zauberman, 2003) in all three sectors, although more search behavior was found within the air travel sector than the others (Johnson et al., 2004). However, households searching within the air travel sector were more likely to gravitate toward a preferred Web site over time (Johnson et al., 2004), thus indicating a propensity for less search in the future. Not surprising, a relationship between activity and depth of search was significant for all sectors indicating households that were more active in a sector were more likely to search across sites (Johnson et al., 2004).

Following in the footsteps of Johnson et al. (2004), Zhang et al. (2006) also examined the search behavior of households, albeit using data collected four years later. The time span between the datasets highlighted the contrasting search behavior of households from the infancy of e-commerce

and message boards" (Danaher et al., 2006, pgs. 186–187).

to its relative maturity. Looking at both product price and the quality of the e-commerce store, an analytical model and propositions of a household's online search behavior were created. Examining two of the three same sectors as Johnson et al. (2004) (music and air travel) and one new sector (computer hardware), linear regression models were used to test hypotheses derived from the analytical model's propositions. The hypotheses sought to determine the relationship of search depth to search cost, product characteristics, previous search behavior, and consumer characteristics.

Compared to prior research, overall search depth increased and loyalty to a Web site decreased (Zhang et al., 2006), which is contrary to the belief that households would gravitate towards a preferred Web site over time (Johnson et al., 2004). It was also found that households took both the price of the product and the quality of the e-commerce store into consideration (Zhang et al., 2006). Like Danaher et al. (2006), who found age was an important moderating variable for visit duration to a site, age was also found to be positively related to search depth, although only within the air travel sector. All told, the linear regression models accounted for 4.4% to 11.5% of the adjusted R^2 (Zhang et al., 2006), indicating other metrics may also be of interest for explaining search depth.

2.2.3 Purchasing

Montgomery et al. (2004) sought to predict purchase conversion by examining the path a visitor took as they browsed a Web site. The path was assumed to provide clues into the goals of the visitor and consisted of the sequence and types of pages (e.g., home, product, and category) viewed throughout a session.

Figure 3 provides an example of two distinct paths from two visitors who eventually arrive at the same product page. The first visitor appears to have taken a direct route to the product of interest, thus exhibiting a deliberate path. In contrast, the second visitor appears to be browsing, due to the number of product and category pages being viewed. These two behaviors are very similar to the search behavior dimension of the shopping strategy typology from Moe (2003). However, unlike Moe (2003) which categorized a visitor as having a static search behavior for the entire session, the dynamic multinomial probit model by Montgomery et al. (2004) included the ability to account for changes to a visitor's search behavior within a session. Therefore, while a visitor may not have a specific goal at the beginning of a session, they may transition at some point in the ses-

sion into having a goal or vice-versa.

Visitor	Session[a]
1	$\langle H, C, P \rangle$
2	$\langle H, C, P, P, P, C, P, P, C, P, P, H, C, P \rangle$

[a] Types of pages: H = Home; C = Category; P = Product.

Figure 3.: Category of Pages Viewed in a Path

One month of panel data focusing only on visitors to BarnesandNoble.com was used to empirically evaluate the accuracy of the proposed model. First, the general accuracy of the model's ability to correctly predict future paths based on prior paths of the same visitor was evaluated. Using a holdout sample, future paths were predicted with 83.2% accuracy (Montgomery et al., 2004). Second, it was found that models which allowed for search behavior to change within a session were more accurate at predicting paths than other models (Montgomery et al., 2004). Lastly, the accuracy of predicting purchase conversion by the end of a session using path information of that session was evaluated. As a path is a discrete set of pages viewed, the purchase conversion prediction can be calculated after each page viewed. Using a holdout sample the accuracy after one page and six pages viewed was 10.4% and 21.2%, respectively, with the accuracy increasing as more pages were viewed (Montgomery et al., 2004).

Predicting if a purchase would be made during a visitor's next visit to a site, Van den Poel and Buckinx (2005) investigated the importance of four different categories of metrics on purchase prediction. The first category of metrics aggregated clickstream measures for a particular visitor regarding all their previous visits to the site. The second category provided detailed clickstream metrics according to the particular content being visited (e.g., a product page), as opposed to the entire site in general. The third and fourth categories dealt with demographic and past purchase metrics, respectively.

An exploratory approach to cull the list of 92 available metrics down to a reasonable set for use in logit models was done via three competing metric selection methods. Using 10 months of data from a commercial wine seller, 11 distinct metrics were used to create the models corresponding to the metric selection method employed. The criteria for judging the models found the best model using the validation dataset was low in accuracy for both the proportional chance criterion

(Morrison, 1969) and the area under receiver operating characteristic curve criterion (Fischer et al., 2003). Although not extremely accurate, Van den Poel and Buckinx (2005) did find the detailed metrics provided the greatest predictive performance.

Padmanabhan et al. (2001) investigated the implications of using incomplete clickstream data to train models for prediction purposes. Specifically, the purpose was to determine if a purchase would occur during the remainder of a session or at any point in the future. The potential problem of using incomplete data is only a visitor's browsing behavior for the particular site of interest is observed. For instance, figure 4 provides an example of two-types of data for two visitors. Examining only the site-centric data, it appears both visitors are similar since they have both visited three pages at site A. However, if user-centric data is examined instead, the picture of the two visitors' browsing sessions is much different. Visitor 1 is visiting two other sites in addition to site A, whereas visitor 2 is only visiting site A.

Visitor	User-centric[a]	Site-centric[b]
1	$\langle A_1, A_2, B_1, A_3, C_1, C_2 \rangle$	$\langle A_1, A_2, A_3 \rangle$
2	$\langle A_1, A_2, A_3, \rangle$	$\langle A_1, A_2, A_3 \rangle$

[a] Notation: X_y indicates the y^{th} page viewed from site X.
[b] Assumes site-centric data is for site A.

Figure 4.: User-centric Versus Site-centric Data

To explore the effects of using such incomplete data, Padmanabhan et al. (2001) recreated a site-centric dataset from six months of user-centric data. A linear regression, logistic regression, classification tree, and neural network were created for each class of problem (i.e., purchase in the current session or future purchase). The results of each model were compared against the site- and user-centric datasets. All the models using user-centric data had significantly higher lifts compared to the models built using site-centric data (Padmanabhan et al., 2001). In addition, some metrics found to be significant in site-centric models were insignificant in the user-centric models, thus leading to the possibility that erroneous conclusions may be reached from relying solely on site-centric data (Padmanabhan et al., 2001). Lastly, some highly significant metrics were only available in the user-centric dataset (Padmanabhan et al., 2001) highlighting the importance of using a complete picture of a visitor's browsing behavior.

Instead of attempting to predict the probability of purchasing as a discrete purchase or not-purchase outcome, Sismeiro and Bucklin (2004) viewed the purchasing process as a series of tasks to be completed. Each task (e.g., find a product, add the product to the shopping cart, checkout) is sequential in nature and requires prior tasks to already have been completed. Therefore, the product of a chain of conditional probabilities can be calculated for a visitor after each task has been completed (Sismeiro and Bucklin, 2004). For example, the probability can be calculated for a visitor adding a product to their shopping cart given the visitor has already found the product. In addition to the task-completion aspect of their model, Sismeiro and Bucklin (2004) also allowed for heterogeneity across visitors at the geographical county level.

In order to evaluate the multi-stage binary choice model, Sismeiro and Bucklin (2004) gathered 70 days of clickstream data from a major commercial Web site in the automotive industry. Three sequential tasks were defined as critical junctures leading up to the purchase of an automobile: completing the configuration of an automobile; inputting personal information; and completing an order. To determine the effectiveness of the task-completion approach, two single-task hierarchical probit models, one with dummy variables representing the completion of the first two tasks and one without, were compared against the multi-stage binary choice model. The multi-stage model outperformed both single-task models in hit rate and mean square error (MSE) for predicting vehicle orders (Sismeiro and Bucklin, 2004). In addition, the multi-stage model demonstrated that some metrics' effect signs differ depending on the task and some metrics are valuable for predicting the completion of some tasks but not for others (Sismeiro and Bucklin, 2004). One stated limitation of Sismeiro and Bucklin (2004) was the requirement that each task must be performed in the order specified. The model cannot consider alternate routes a visitor takes at a site that may also lead to a purchase[6].

Recognizing that visitors may have distinct purchasing patterns, Moe and Fader (2004b) investigated how purchasing probabilities can be improved by taking into account visitor heterogeneity and visit history. Specifically, they created a conversion model which contained six components. The first component was a baseline probability of purchasing for each visitor which was independent of the visitor's past history. The positive effect on purchasing (i.e., visit effects) was the second component and assumed that each visit increased, although by varying amounts, the likelihood

[6]Sismeiro and Bucklin (2004) cite Amazon.com's "One-Click" checkout service as an example of an alternate route.

of a future purchase. The third component, purchasing threshold, allowed for a negative effect on purchasing which may be caused by the risk-adverseness or reluctance of a visitor to purchase. A decreasing threshold would indicate less of a negative effect on a visitor's purchasing probability. The fourth component permitted heterogeneity across visitors by differing visit effects and purchasing thresholds for each visitor. The fifth component allowed for changes and evolutions in the visit effects and purchasing threshold over time. Lastly, the model included a component to remove shoppers who were considered "hard-core never-buyers" and had no intention of ever purchasing (Moe and Fader, 2004b).

Using an eight-month sample of panel data focusing on Amazon.com, visit effects were found to accumulate over time and increase purchasing probabilities (Moe and Fader, 2004b). However, the conversion model showed an overall decrease in purchasing probabilities over time. The third and fifth components showed repeat visits had less of an impact on purchasing over time and purchasing thresholds increased as visitors became more experienced (Moe and Fader, 2004b). These results mirror Moe and Fader (2004a) indicating the importance of exploring visitor-level data as the conversion model contradicted the aggregated dataset's demonstration of increasing purchasing probabilities. Lastly, the conversion model outperformed a logistic regression model, duration model, beta-binomial, and historical conversion rates by having the lowest relative error in predicting conversion rates (14.7% predicted versus 15.7% actual).

2.2.4 Goal Achievement

Chatterjee et al. (2003) analytically modeled a visitor's response to banner advertisement exposure on an ad-sponsored magazine Web site[7]. The proposed model allowed for heterogeneity of visitors and their sessions (both within and across) to the Web site. Three benchmark models, with varying levels of heterogeneity, were used to compare the proposed model.

Using data from 1995, the ability of the model to predict the click-through of banner advertisements of 3,611 visitors for two sponsors was tested. The proposed model obtained a 41% hit rate compared to the three alternative models obtaining a 2.4%, 24.2%, and 33.3% hit rate, respectively (Chatterjee et al., 2003). As expected it was found that "wearout" to banner exposure was a factor

[7] A notable point about this type of goal achievement is it can occur multiple times within the same session (i.e., a visitor can click multiple banner advertisements during a session). Although multiple purchases or other goals may be achieved within the same session, it is unlikely to occur with much frequency.

and thus the probability of clicking on a banner advertisement was higher when a visitor was first exposed to the advertisement. In addition, the "wearout" concept also extended over sessions such that infrequent visitors were more likely to click on a banner advertisement than frequent visitors (Chatterjee et al., 2003).

2.3 Datasets

Table 2 provides information about the dataset used in each study. As clickstream research is typically data-driven, the results found in prior literature may be specific to particular datasets. Thus, having knowledge of the datasets used may be helpful in understanding differing results. Each dataset is broken down by type, sector the site or sites of interest belong to, year when the data was collected, duration of data collection, and size of the dataset.

2.3.1 Type

The type of dataset used can be categorized as either site-centric or user-centric, terms coined by Padmanabhan et al. (2001) to refer to the focus of a clickstream dataset. Site-centric data is focused on the site itself and is defined as "…clickstream data collected at a site augmented with user demographics and cookies to identify users" (Padmanabhan et al., 2001, pg. 154). Although site-centric data is advantageous in terms of being readily available to site-owners (although they might not have access to demographics) and including all traffic to a site, it can only provide information on a visitor's browsing behavior on that site. A visitor's entire browsing session (i.e., user session), which may include browsing at other sites, cannot be obtained from site-centric data.

User-centric data overcomes the disadvantage of site-centric data by providing an entire visitor's session regardless of the different sites visited. Having complete information, user-centric data has proven to be more accurate than site-centric data when building models predicting purchasing probabilities (Padmanabhan et al., 2001). User-centric data is defined as "…site-centric data plus data on where else the user went in the current session" (Padmanabhan et al., 2001, pgs. 154–155). User-centric data is typically obtained from randomly selected participants who are representative of the population at large. Unfortunately, some more recent user-centric datasets lack details of each page a user visits during their session. This limitation restricts the examination of paths and other such techniques from being studied using these datasets.

In addition to the "pure" site- and user-centric datasets, some studies construct site-centric datasets from user-centric data for a single site or set of sites. These constructed site-centric datasets typically view the site or sites of interest in isolation without regard to where a visitor may browse at other sites (e.g., Montgomery et al., 2004).

2.3.2 Sector

The sites examined in a study's dataset can be categorized into a general sector according to the main purpose of the site or type of products sold. Awareness of the sector may be desirable since browsing behavior has been found to vary by sector (Johnson et al., 2004). Therefore, the results obtained from a site in one sector may not be generalizable to sites in another sector. For example, Van den Poel and Buckinx (2005) looked at an e-commerce site selling wine, which is within the food sector. The results from such a sector may not generalize well to other sectors as wine is a perishable good which may require restocking as consumed. A site within the consumer electronics sector may have drastically different traffic patterns and factors that lead to purchase due to the non-perishable aspect of the products sold.

Sites were categorized into sectors according to the Open Directory Project (ODP). A search on a site's domain was done on ODP and the most relevant category returned was used as the sector. For studies which did not explicitly mention the sites used, the sector was located according to the general purpose or type of product sold. As sites may change drastically over time the purpose or type of product sold at the dataset's time was used to categorize the site[8].

2.3.3 Time, Duration, and Size

As the Internet has seen an evolution of visitors' behavior to sites (Zhang et al., 2006), the year or years in which a dataset was collected can have a direct impact on the results obtained (cf. Johnson et al., 2004; Zhang et al., 2006). The duration of data collection can also have profound results. For instance, collecting data for one month in a cyclical industry may result in differing conclusions when compared to data collected in the same industry over a longer period of time. Lastly, the size of the dataset is provided. For site-centric data the number of monthly visitors can be con-

[8] Amazon.com circa 1998 sold only books (Moe and Fader, 2004b) and thus would be assigned to the Book sector. Today, however, Amazon.com sells many different categories of products ranging from consumer electronics to bedding and thus would be assigned to the General sector.

sidered an accurate, albeit conservative, estimate of actual visitors[9]. As user-centric data represent only a sub-sample of all visitors to a site, such datasets cannot accurately represent the size of a site. However, many studies using user-centric datasets focus on large well-known sites such as Amazon.com (Moe and Fader, 2004b), BarnesAndNoble.com (Montgomery et al., 2004), and CD-Now.com (Moe and Fader, 2004a) in order to obtain adequate sized samples and generalizability.

[9] As part of the preprocessing of clickstream data, sessions consisting of only a single page are typically discarded (Bucklin and Sismeiro, 2003). Therefore, site-centric data measures of monthly visitors are likely a conservative estimate of actual unique visitors to a site.

Table 2: Prior Literature: Datasets

Article	Dataset				
	Type[a]	Sector	Year	Duration	Monthly Visitors[b]
MULTIPLE OBJECTIVES					
Kalczynski et al. (2006)[f]	site-centric	construction, financial, government, insurance, & travel	N/A	N/A	97
Moe (2003)	site-centric	nutritional supplements	2000	7 weeks	3,508
BROWSING					
Bucklin and Sismeiro (2003)	site-centric	autos	1999	1 month	164,429
Danaher et al. (2006)	user-centric	all	2000	1 month	1,665[d]
Johnson et al. (2004)	user-centric	books, music, & travel	1997-1998	1 year	893[d]
Moe and Fader (2004a)	site-centric	books & music	1998	8 months	741[d]
Park and Fader (2004)	user-centric	books & music	1997-1998	8 months	1,039[d]
Zhang et al. (2006)	user-centric	music, computer hardware, & travel	2002	6 months	2,277[d]
PURCHASING					
Moe and Fader (2004b)	site-centric[c]	books	1998	8 months	536
Montgomery et al. (2004)	site-centric[c]	general	2002	1 month	1,160
Padmanabhan et al. (2001)	both[c]	multiple	N/A	6 months	3,297

Continued on Next Page…

Table 2: Prior Literature Datasets – Continued

Article	Dataset				
	Type[a]	Sector	Year	Duration	Monthly Visitors[b]
Sismeiro and Bucklin (2004)	site-centric	autos	2000-2001	70 days	37,597
Van den Poel and Buckinx (2005)	site-centric	food	2001-2002	11 months	126
GOAL ACHIEVEMENT					
Chatterjee et al. (2003)	site-centric	magazine	1995	7 months	479[e]

[a] A dataset is considered site-centric if only information about the particular site or sites under study were examined independently of any other sites the visitor may have visited. Thus, while the data may be user-centric in nature (i.e., panel data) the researchers have taken a site-centric approach by disregarding browsing behavior at other sites.

[b] Monthly visitors were calculated as shown in equation 2.1. Traffic patterns were assumed to be constant throughout a dataset's duration. If the specific dates of the dataset were not provided approximations were made.

$$\left(\frac{\text{total number of visitors}}{\text{duration of dataset in days}}\right) \times 30 \tag{2.1}$$

[c] Indicates a dataset constructed from user-centric data. These datasets only reflect a subset of all monthly visitors to a site.

[d] Indicates total monthly visitors across all sites analyzed in the dataset.

[e] Indicates a subset of all monthly visitors to a site that met specified criteria.

[f] Captured clickstream data from experimental subjects performing a prespecified task.

2.4 Metrics

Table 3 provides information about the metrics used in each study. Categorized by the focus of the research, the metrics used are described in terms of their level of analysis and general type.

2.4.1 Analysis Level

The metrics used for clickstream research can be defined at four basic levels of analysis: session, site, sector, and user. At the session-level, which is the most detailed, each session of a visitor is typically treated as an independent datapoint. A session-level metric is based only on information within the same session (e.g., number of clicks in a session, time spent during a session). When all sessions at a particular site for a visitor are aggregated together, they represent the site level of analysis. At the site level of analysis each visitor is considered a datapoint, regardless of the number of sessions at a site. Metrics at the site-level can provide a historical perspective of a visitor's browsing history at that site (e.g., conversion rate for the visitor).

The sector-level performs another level of aggregation for all sites visited within the same sector. As more than one site is available, the sector-level can provide metrics that compare a visitor's browsing and purchasing habits across various sites within the sector (e.g., percentage of visits to site A). Finally, the user-level aggregates every sector, which includes all sites and all sessions, of a visitor's browsing behavior together. Similar to sector-level metrics, user-level metrics are also able to compare browsing and purchasing habits, albeit at a higher level of analysis (i.e., the sector)[10].

Figure 5 is an example of the different levels of analysis possible for a single user. The user U_1 depicted in figure 5 had ten sessions (S_{1-10}) at four sites (I_{1-4}) within three sectors (C_{1-3}). Taking a site-centric approach, either session-level or site-level metrics could be used. For instance, a site-centric approach at site I_1 could examine each of the user's sessions ($S_{1-3,5}$) as individual datapoints; all sessions aggregated to the site-level (I_1) as a single datapoint; or a combination of both where the last session (S_5) is used at the session-level and all previous sessions (S_{1-3}) are aggregated at the site-level for historical metrics.

Taking a user-centric approach, not only can session- and site-level metrics be used, but also

[10] Catledge and Pitkow (1995) also noted varying levels of analysis. However, they only considered the session and user level of analysis.

User	(U)	U_1=$\{C_{1-3}\}$			
Sector	(C)	C_1=$\{I_{1-2}\}$		C_2=$\{I_3\}$	C_3=$\{I_4\}$
Site	(I)	I_1=$\{S_{1-3,5}\}$	I_2=$\{S_{4,9}\}$	I_3=$\{S_6\}$	I_4=$\{S_{7-8,10}\}$
Session	(S)	S_1, S_2, S_3, S_5	S_4, S_9	S_6	S_7, S_8, S_{10}

Figure 5.: Example Metric Level of Analysis

sector- and user-level metrics. For instance, continuing the example of user U_1 from figure 5, at sector C_1 all site-level data can be aggregated as a single datapoint for the user. In addition, the site- and session-level metrics would also be available for both sites (I_{1-2}) and the corresponding sessions ($S_{1-5,9}$). At the most general level, user metrics aggregated from all sectors (C_{1-3}) for the user would be represented as a single datapoint. Furthermore, all other level-of-analysis metrics would be available for all the sectors, sites, and sessions of the user.

Each level of analysis can be further broken down into general and detailed metrics. General metrics refer to any metrics at a particular level of analysis that includes all relevant behavior, without regard to the underlying content of the site. Detailed, on the other hand, breaks metrics down by the type of content being viewed (Van den Poel and Buckinx, 2005). For instance, a general session-level metric would be the number of pages viewed in a session whereas a detailed session-level metric would be the number of product pages viewed in a session. Aggregating all the detailed metrics for a level of analysis would result in the general metrics for the same level of analysis.

2.4.2 Metric Categories

Direct marketers have used Recency, Frequency, and Monetary (RFM) metrics to segment customers for decades (Shaver, 1996; Stone and Jacobs, 2001) and summarize their prior behavior (Fader et al., 2005). Recency refers to the amount of time elapsed since a particular action or behavior has been observed, frequency is concerned with the number of times the same action or behavior is made, and monetary deals with the amount of money spent on current or past purchases. For example, the amount of time since a visitor last visited a Web site, the number of pages viewed during a session, and the total amount spent on a previous purchase, would represent a recency, frequency, and monetary metric, respectively.

As the basic underlying goal of identifying valuable customers is common in both direct marketing and clickstream research, the classic RFM metrics can be a logical starting point for categorizing the metrics used in the clickstream literature. However, differences between the online and offline environment affect the data available and thus the types of metrics which can be used[11]. Within the RFM metrics both recency and frequency are well represented in clickstream research, but monetary metrics have not seen widespread usage since datasets do not explicitly contain pricing information of visitors' purchases[12].

Besides RFM metrics, user characteristics (i.e., demographics) have been used in both direct marketing and clickstream research with some regularity. Although not available in a user's clickstream, user-centric panel data typically obtain demographic data separately which are then mapped to the appropriate clickstream. Duration and timing are two measures more specific to clickstream research, due to the ease at which they can be obtained. Duration deals with the amount of time spent doing an action or behavior, whereas timing is the rate at which an action or behavior is done. The amount of time spent on a Web site and the visitation rate of a visitor to a site are examples of duration and timing, respectively. Lastly, the structure of the Internet and characteristics of Web sites and their pages lend themselves to a wide variety of other metrics which do not fit into the previously discussed metrics. The referring Web site, number of links on a page, and size of a page in bytes are all examples of the type of metrics which belong in the "other" category.

Table 3 classifies the metrics used in prior literature according to the categories they belong: demographics, recency, frequency, monetary, duration, timing, and other metrics.

[11] Take the example of determining the number of products viewed for an online store versus an offline catalog. Online a simple count of the number of pages viewed with product information would provide the relevant information. Offline attempting to gain information for such a metric would be prohibitively expensive since some type of observation would be needed for each viewer of the catalog.

[12] Monetary metrics can be obtained from secondary sources and has been examined in Van den Poel and Buckinx (2005), but it is not a natural byproduct found in server logs and other such sources that clickstream data are typically gathered from. Some user-centric datasets; however, do provide monetary values for products sold.

Table 3: Prior Literature: Metrics

Article	Analysis Level	Demographics	Recency	Frequency	Monetary	Duration	Timing	Other
MULTIPLE OBJECTIVES								
Kalczynski et al. (2006)	session							Y
Moe (2003)	session			Y		Y		Y
BROWSING								
Bucklin and Sismeiro (2003)	session & site			Y		Y		Y
Danaher et al. (2006)	session	Y		Y		Y		Y
Johnson et al. (2004)	sector			Y				Y
Moe and Fader (2004a)	session			Y			Y	
Park and Fader (2004)	sector			Y			Y	
Zhang et al. (2006)	site & sector	Y		Y	Y			
PURCHASING								
Moe and Fader (2004b)	session			Y				
Montgomery et al. (2004)	session	Y						Y
Padmanabhan et al. (2001)	site & sector	Y		Y		Y		Y
Sismeiro and Bucklin (2004)	session & site	Y		Y		Y		Y
Van den Poel and Buckinx (2005)	session & site	Y	Y	Y				Y

Continued on Next Page…

Table 3: Prior Literature Metrics – Continued

Article	Analysis Level	Demographics	Recency	Frequency	Monetary	Duration	Timing	Other
GOAL ACHIEVEMENT								
Chatterjee et al. (2003)	session & site		Y	Y		Y		

2.5 Conclusion

The preceding sections in this chapter organized and summarized research using clickstream data for prediction. All told, the majority of research has focused on the browsing (Johnson et al., 2004) and purchasing behavior (Padmanabhan et al., 2001) of Web users at e-commerce sites, with little attention being paid to alternative objectives (i.e., goal achievement) or contexts (e.g., informational Web sites). As there are many different types of Web sites (Jaillet, 2003), focusing on just e-commerce sites is done at the expense of understanding visitor behavior at other interesting and valuable non e-commerce Web sites.

In terms of data, user-centric datasets are commonly used when examining browsing behavior, but with the exception of Padmanabhan et al. (2001) is non-existent for purchasing or goal achievement behaviors. The sectors examined for browsing behavior generally overlap (e.g., books and music) allowing comparisons of results. For purchasing and goal achievement, however, there is little overlap and some of the sectors analyzed differ substantially from one another (e.g., automobiles versus wine). According to Zhang et al. (2006), who found browsing differs by sector, such little overlap may make results difficult to compare over studies. Lastly, many of the datasets are fairly dated. Although beneficial from the standpoint of comparing and contrasting changes over time (cf. Johnson et al., 2004; Zhang et al., 2006), the vast changes in the Internet and Web users over the past few years may point toward a need for more recent and thus relevant datasets.

Although many studies used metrics that did not fit neatly into the categories of table 3, general patterns of the types of metrics used can still be seen. Overall, frequency appears to be the most commonly used type of metric as every single study except for Kalczynski et al. (2006) and Montgomery et al. (2004) included some aspect of counting in their models. Duration metrics were also commonly used for all types of research. Lastly, timing metrics were more heavily used in browsing while recency was more common in purchasing and goal achievement. Determining how well these types of metrics do for other objectives and contexts along with finding a common set of metrics can provide the basis for better understanding visitor behavior. Furthermore, looking outside these metric types into the "other" category[13] can also help provide explanation into the "whys" of visitor behavior.

[13]However, these "other" metrics should be readily available to all Web sites and not be an artifact of a particular site or how it is organized.

Chapter 3

Theory

Information Foraging Theory (IFT) "... aims to explain and predict how people will best shape themselves for their information environments and how information environments can best be shaped by people" (Pirolli, 2007, pg. 3). Simply stated one aspect of IFT is its ability to explain the behavior of a person as they search for information within a pliable environment. Central to the theory of information foraging are the concepts of information scent and patches. Information scent is the driving force of why a person makes a navigational selection amongst a group of competing options. Information patches are distinct areas of the search environment which differ in their informational content. The synthesis of behavior (i.e., information scent) and environment (i.e., information patches) provides for a rich theory of information foraging.

IFT has a strong theoretical foundation by drawing upon Optimal Foraging Theory (OFT) (Stephens and Krebs, 1986) and the Adaptive Control of Thought-Rational Theory (ACT-R) (Anderson et al., 2004), two well-known theories within their respective fields. OFT is an ecological theory concerned with explaining the foraging behavior of animals as they hunt for food. ACT-R is a psychological theory of the human mind that includes the cognitive architecture and process by which cognition works. Within IFT, OFT is used to explain the behavioral elements of people foraging for information (i.e., why they go about searching), whereas ACT-R's purpose is to explain the mechanism of how the behavior is being driven at the cognitive level.

The remainder of this chapter is organized as follows. First an introduction of OFT and ACT-R are provided in §3.1 and §3.2 as background information for IFT. Then details regarding the two central concepts of IFT are presented in §3.3, followed by two versions of a model that test the theory.

3.1 Optimal Foraging Theory

The aim of optimal foraging theory (OFT) is to explain the feeding behaviors and adaptations of animals (Stephens and Krebs, 1986). OFT has been used to describe the commuting behavior of seabirds to distal feeding grounds (Nevitt, 2000); assess the nutritional ratios of ants' food (Kay, 2002); predict the feeding strategies of coyotes (MacCracken and Hansen, 1987); and test the group foraging behaviors of cranes (Alonso et al., 1995). In addition, OFT has also been applied to humans by explaining the hunting and gathering practices of the Aché of eastern Paraguay (Hawkes et al., 1982); variability in Amazonian Indians' diet selections (Hames and Vickers, 1982)[1]; decisions between ambiguous and unambiguous choices (Rode et al., 1999); and applicability of OFT in information seeking behaviors (Sandstrom, 1994).

The overarching assumption of OFT is animals have developed beneficial foraging adaptations and behaviors that increase their net energy. A gain in net energy (above an animal's metabolic requirement) allows spare energy to be spent on vital non-feeding activities such as fighting, fleeing, and reproducing (Stephens and Krebs, 1986). As animals with higher levels of spare energy are more likely to survive and reproduce, successive generations are assumed to inherit those beneficial foraging adaptations and behaviors.

Following MacArthur and Pianka (1966), the general concepts used in OFT are that of predators, prey, and patches (Stephens and Krebs, 1986). Predators are the animals doing the foraging (i.e., hunting for food) and their behaviors are the focal point of OFT. Prey refers to any item of food that a predator may consume such as a rabbit, berry, or plant root. Each type of prey differs in their prevalence in the environment along with the amount of energy the predator expends and gains from chasing and eating the prey, respectively. A patch is some area of the environment which contains prey. Like prey, patches of different types demonstrate variability in terms of the net energy a predator gains from foraging within them.

Predators are assumed to forage for food according to a sequential search–encounter–decision process (Stephens and Krebs, 1986). While searching, an animal uses its sensory abilities to pick up on cues to help locate prey or patches. For example, seabirds use their sense of smell to (1) locate patches over thousands of kilometers from their nesting colony and (2) find prey within those patches (Nevitt, 2000). Without sensory guidance, the forager's probability of encountering prey

[1] A more complete review of OFT's use in anthropological research can be found in Smith (1983).

or patches is effectively reduced to random chance. Searching stops once prey or a patch has been located (i.e., an encounter has occurred). At the point of encounter the forager makes a decision of how to proceed.

The two conventional models of OFT agree on the search–encounter–decision process, but fundamentally differ in what decision to make once an encounter takes place. The prey model (Charnov and Orians, 1973) asks the question "attack or continue searching?" (Stephens and Krebs, 1986, pg. 13) when encountering prey. In the patch model (Charnov, 1976) the forager asks the question "how long to stay in a patch?" (Stephens and Krebs, 1986, pg. 14) when a patch has been encountered.

As the overarching assumption of OFT is the increase of net energy, both models are concerned with maximizing the average rate of energy intake. Therefore, each model uses a variant of Holling's disc equation (equation 3.1) (Holling, 1959). The average rate of energy intake is represented as R and is what both the prey and patch models are maximizing. Depending on the model used λ is either the rate of encounter of prey or a patch. $\bar{e}$ is the average energy gained from each encounter, s is the search cost per unit of time, and finally $\bar{h}$ is the average handling time per encounter. Within each model the theory assumes predators have perfect information regarding the characteristics (e.g., $\lambda, \bar{e}, \bar{h}$) of its prey or patches (Stephens and Krebs, 1986). Even though animals do not possess perfect information, the theory has still found empirical support even when the assumption of perfect information has been violated (Kay, 2002).

$$R = \frac{\lambda\bar{e} - s}{1 + \lambda\bar{h}} \tag{3.1}$$

3.1.1 Prey Model

The prey model determines if an animal should attack and consume a particular type of prey or continue searching for other prey types (Charnov and Orians, 1973). The decision is made by maximizing the average rate of energy intake by prey type to find the optimal diet (Stephens and Krebs, 1986). Within the prey model there are n different prey types encountered at random with i representing the i^{th} prey type. Let D represent the set of prey types such that $D = \{1, 2, \ldots, n\}$. Associated with each prey type are the following characteristics[2]:

[2]Notations follow Pirolli (2007).

- t_{Bi} = average time between locating prey of type i.
- λ_i = rate of encounter of prey type i when searching ($1/t_{Bi}$).
- t_{Wi} = handling time associated with pursuing, capturing, and consuming prey type i.
- g_i = net energy gain from consuming prey type i.
- π_i = profitability of prey type i (g_i/t_{Wi}).
- p_i = probability of attacking prey type i upon encounter[3].

The long-term average rate of energy intake for all prey types is determined from equation 3.2 (a variant of Holling's disc equation) (Stephens and Krebs, 1986).

$$R = \frac{\sum_{i \in D} p_i \lambda_i g_i}{1 + \sum_{i \in D} p_i \lambda_i t_{Wi}} \tag{3.2}$$

The inclusion of some prey types into a forager's diet, when compared to the alternatives, may never be worth the energy to attack and consume. Determining which prey types should be excluded from consideration is expressed via the probability of attacking a prey type (p_i). In order to maximize R the prey model follows the zero-one rule which states a prey type is either always attacked ($p_i = 1$) or always ignored ($p_i = 0$) (Stephens and Krebs, 1986). As expressed in equation 3.3, prey types are excluded when the profitability of prey type i is less than the average rate of energy intake of all other prey types[4].

$$p_i = \begin{cases} 0 & \text{if } \pi_i < \frac{\sum_{j \in D-\{i\}} \lambda_j g_j}{1+\sum_{j \in D-\{i\}} \lambda_j t_{Wj}} \\ 1 & \text{otherwise} \end{cases} \tag{3.3}$$

Once the probability of attacking each prey type has been determined, the decision then turns to selecting which prey types to include for an optimal diet. The two-step prey algorithm makes the selection based on the profitability of a prey type compared to the current average rate of energy (R) (Stephens and Krebs, 1986). The first step is to rank the k remaining prey types in order of decreasing profitability such that $\pi_1 > \pi_2 > \ldots > \pi_k$. In the second step each prey type is added to the forager's diet until equation 3.4 is true. The last prey type added to the diet is the lowest

[3]Probability is the only characteristic of a prey type the forager has control over.

[4]The derivation of equation 3.2 to determine which prey types should or should not be attacked when encountered (i.e., equation 3.3) can be found in Stephens and Krebs (1986).

ranking prey type in the diet. If the equation is never true then all prey types are included in the diet.

$$R(k) = \frac{\sum_{i=1}^{k} \lambda_i g_i}{1 + \sum_{i=1}^{k} \lambda_i t_{Wi}} > \pi_{k+1} \tag{3.4}$$

Figure 6 shows a simulated example of ten different prey types available to a predator. The figure illustrates the relationship between the profitability of a prey type (π_k) against the average rate of energy ($R(k)$). In this example, the average rate of energy is maximized when the five most profitable prey types are added to the forager's diet. This maximization is the point at which the current rate of energy ($R(5) = 0.9888$) is greater than the profitability of the next prey type ($\pi_6 = 0.9838$) (equation 3.4). As illustrated, adding additional prey types or removing any of the five selected prey types leads to a decrease in R and thus a sub-optimal diet.

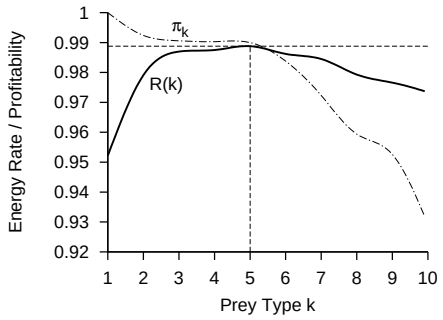


Figure 6.: OFT: Simulated Optimal Diet

Notable about equation 3.4 is that the inclusion of a prey type into a forager's diet is independent of its rate of encounter (Charnov and Orians, 1973). The decision to add a prey type is dependent, however, on the rate of encounter for those prey types ranked higher than the current prey type. For example, in a situation with two prey types the decision to include the second prey type is only dependent on (1) its profitability (π_2) and (2) the rate of encounter, net energy, and handling time of the first prey type (λ_1, e_1, h_1).

Foraging Example

To illustrate the prey model consider a hypothetical example of a brown bear foraging for food[5]. Brown bears are known for their diverse diets (Garshelis, 2007) and therefore the decision of which prey types to include in their diets is germane to the discussion of the prey model. Assume that four different prey types are present in the bear's environment. Each of the four prey types and their characteristics are listed in table 4.

Table 4: OFT: Example Prey Types for a Brown Bear

Prey Type	t_B	t_W	g	π	p[a]
Deer	$3{,}600\,\mathrm{sec}$	$2{,}580\,\mathrm{sec}$	$3{,}200\,\mathrm{kCal}$	$1.2403\,\mathrm{kCal/sec}$	1
Berries	$12\,\mathrm{sec}$	$180\,\mathrm{sec}$	$200\,\mathrm{kCal}$	$1.1111\,\mathrm{kCal/sec}$	1
Squirrels	$6\,\mathrm{sec}$	$600\,\mathrm{sec}$	$610\,\mathrm{kCal}$	$1.0167\,\mathrm{kCal/sec}$	1
Chipmunks	$6\,\mathrm{sec}$	$720\,\mathrm{sec}$	$500\,\mathrm{kCal}$	$0.6944\,\mathrm{kCal/sec}$	0

[a] As calculated from equation 3.3.

Based on the characteristics of the other prey types, the profitability of chipmunks will never be high enough to warrant the bear eating them and therefore $p_{chipmunks}$ is set to 0 (equation 3.3). For the first step of the prey algorithm the remaining prey types are ranked in descending order according to their profitability which yields $\pi_{deer} > \pi_{berries} > \pi_{squirrels}$. The results of the iterative second step of the prey algorithm can be seen in table 5. The R column is the long-term average rate of energy for the included prey types (left-hand side of equation 3.4) and the π column is the profitability of the next lowest ranking prey type (right-hand side of equation 3.4). The final column *Stop?* is set to yes if the last added prey type causes the inequality to be true (i.e., $R > \pi$) and set to no otherwise[6].

As seen in table 5 a diet consisting of deer and berries is optimal for the bear. Eating only deer or choosing to eat all three prey types would result in a sub-optimal rate of energy as illustrated by the lower values of R.

[5] The foraging example was adapted from Pirolli (2007).

[6] Although the algorithm would stop after deer and berries are included in the diet, the calculation for including squirrels into the diet was done for illustrative purposes.

Table 5: OFT: Example Diet for a Brown Bear

Included Prey Types	$R(k)$	π_{k+1}	Stop?
Deer	$0.5178\,\mathrm{kCal/sec}$	$1.1111\,\mathrm{kCal/sec}$	No
Deer & Berries	$1.0502\,\mathrm{kCal/sec}$	$1.0167\,\mathrm{kCal/sec}$	Yes
Deer, Berries, & Squirrels	$1.0215\,\mathrm{kCal/sec}$	n/a	n/a

3.1.2 Patch Model

The patch model determines the optimal duration an animal should forage within any number of patch types (Charnov, 1976). The decision of how long to spend in a patch of a particular type is determined by maximizing the average rate of energy intake (similar to the prey model) (Stephens and Krebs, 1986). Within the patch model there are n different patch types with i representing the i^{th} patch type. Let P represent the set of patch types such that $P = \{1, 2, \ldots, n\}$. Associated with each patch type are the following characteristics[7]:

- t_{Bi} = average time between locating patches of type i.
- λ_i = rate of encounter of patch type i when searching ($1/t_{Bi}$).
- t_{Wi} = the amount of time spent searching within patch type i (i.e., patch residence time)[8].
- $g_i(t_{Wi})$ = net energy gain from patch type i when t_{Wi} time units are spent foraging within the patch (i.e., gain function).

The long-term average rate of energy intake for all patch types is determined from equation 3.5 (a variant of Holling's disc equation) (Stephens and Krebs, 1986).

$$R = \frac{\sum_{i \in P} \lambda_i g_i(t_{Wi})}{1 + \sum_{i \in P} \lambda_i t_{Wi}} \tag{3.5}$$

To better illustrate patches and their characteristics consider a hypothetical example of a seabird foraging for food in an environment with multiple patches of a single patch type. Figure 7 details the environment where the solid line represents the seabird's path as it forages. The squares are patches and within the patches are sources of food shown as fish. The horizontal axis represents time, where the time spent within a patch is t_W and the time spent between patches is t_B. As seen

[7] Notations follow Pirolli (2007).

[8] Patch residence time is the only characteristic of a patch type the forager has control over.

in figure 7 the seabird only eats some of the available fish in each of the patches. Since there is a finite amount of food, each patch demonstrates diminishing returns of energy as a function of time. Due to this diminishing return it would have been suboptimal for the seabird to remain in a patch until total depletion.

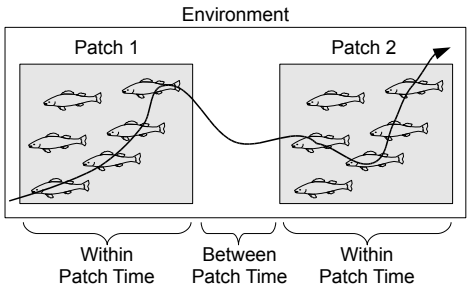


Figure 7.: OFT: Patchy Environment – adapted from Pirolli (2007, pg. 32)

The gain function associated with a patch type, $g_i(t_{Wi})$, determines the amount of energy gained per unit of time spent foraging within a patch. Each gain function is "... assumed to be a well-defined, continuous, deterministic, and negatively accelerated (curving down) function" (Stephens and Krebs, 1986, pg. 25)[9]. Figure 8 illustrates a gain function with the horizontal axis representing time spent foraging in the patch and the vertical axis representing net energy gain. Such a gain function helps explain the behavior of the seabird. Initially the seabird realized a rapid energy gain as there were many fish within the patch. However, as fewer fish were available less energy was gained per unit of time. Thus at some point it was more worthwhile for the seabird to travel to another patch rather than remain in the current patch.

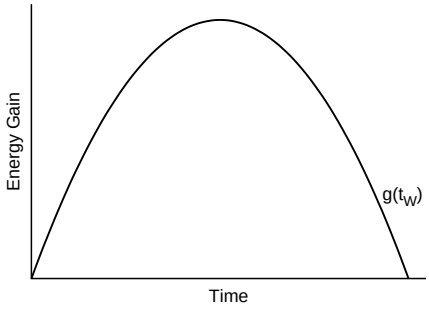


Figure 8.: OFT: Example Patch Gain Function

[9]Stephens and Krebs (1986) acknowledged some gain functions may not exhibit an eventual negative acceleration. When patches are searched systematically their gain functions may exhibit a depletion of resources without any depression.

The determination of how much time to spend in each patch types is made on the basis of the Marginal Value Theorem (Charnov, 1976). For patch types exhibiting a negatively accelerated gain function (as shown in figure 8), the theorem is capable of determining the optimal allocation of time across any number of patch types so as to maximize the average rate of energy within the environment. To obtain optimality the theorem states the predator should continue to forage within a patch type until the marginal rate (i.e., slope of the gain function) equals the average rate of energy gain (equation 3.5). Following such a requirement means by definition the marginal rate of each patch type will be equal to the average rate. Equation 3.6 states the equality condition where $g'(\widehat{t_{Wi}})$ is the marginal rate of patch type i and $R(\widehat{t_{W1}}, \widehat{t_{W2}}, \ldots, \widehat{t_{Wn}})$ is the average rate of energy calculated from the optimal vector of times for each patch type.

$$\begin{aligned} g'(\widehat{t_{W1}}) &= R(\widehat{t_{W1}}, \widehat{t_{W2}}, \ldots, \widehat{t_{Wn}}) \\ g'(\widehat{t_{W2}}) &= R(\widehat{t_{W1}}, \widehat{t_{W2}}, \ldots, \widehat{t_{Wn}}) \\ &\vdots \\ g'(\widehat{t_{Wn}}) &= R(\widehat{t_{W1}}, \widehat{t_{W2}}, \ldots, \widehat{t_{Wn}}) \end{aligned} \tag{3.6}$$

In situations where only a single patch type exists, the average rate of energy intake can be simplified as shown in equation 3.7.

$$R(t_W) = \frac{\lambda g(t_W)}{1 + \lambda t_W} \tag{3.7}$$

The reduction to only a single patch type also simplifies the marginal value theorem as shown in equation 3.8.

$$g'(\widehat{t_{W1}}) = R(\widehat{t_{W1}}) \tag{3.8}$$

Examples of the patch model being used to find the optimal foraging time when (1) only a single patch type exists and (2) when multiple patch types exist are presented next[10].

[10]The patch examples were adapted from Charnov (1976); Stephens and Krebs (1986); and Pirolli (2007).

Single Patch Type Example

Consider an example of a brown bear foraging for berries over a three-year period. Within the bear's environment there exist multiple patches of a single type representing berry bushes. With each season the characteristics of the patch type changes. Table 6 details the characteristics of the patch type for each of the three years.

Table 6: OFT: Example Single Patch Type for a Brown Bear

Year	t_B	$g(t_W)$	R	t_W[a]
Y1	$10\,\text{sec}$	$-0.8 * {t_W}^2 + 6.5 * t_W$	$0.9593\,\text{kCal/sec}$	$3.4629\,\text{sec}$
Y2	$5\,\text{sec}$	$-0.8 * {t_W}^2 + 6.5 * t_W$	$1.5385\,\text{kCal/sec}$	$3.1009\,\text{sec}$
Y3	$5\,\text{sec}$	$-2.5 * {t_W}^2 + 17.5 * t_W$	$3.7702\,\text{kCal/sec}$	$2.7460\,\text{sec}$

[a] As calculated from equation 3.8.

In the first year the optimal time to spend in a patch was $3.4629\,\text{sec}$. Illustrated graphically in figure 9 the optimal point is where the dashed line with its origin at t_B lies tangential to the gain function.

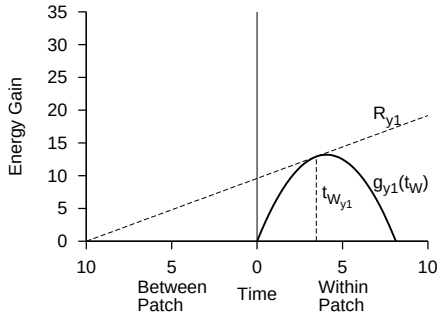


Figure 9.: OFT: Example Year One Patch

In the second year, an area of the environment previously destroyed by wildfire bloomed with berry bushes. This represents an increase in the number of patches available to the bear and therefore a decrease in the time between patches (t_B). Figure 10 shows graphically how the decrease in time between patches leads to a reduction in the time spent within a patch. Although less time is spent per patch, the average rate of energy gain (R_{y2}) is higher during the second year. With lower moving costs, the bear is better served to move to another patch once R_{y2} drops too low.

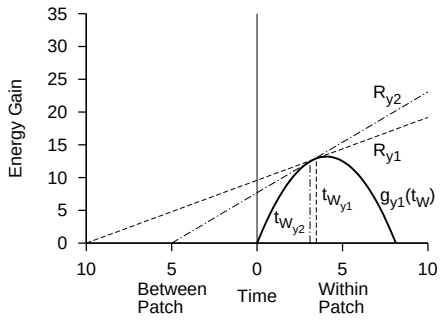


Figure 10.: OFT: Example Year Two Patch

A bountiful rain during the third year increased the density of berry bushes within each patch. A greater density of bushes represents a more valuable patch. Therefore, the gain function for the patch type is changed to reflect greater energy gains per unit of time spent in a patch. Figure 11 illustrates the difference when the gain function of a patch type changes. In this situation the gain function reflected an increase in energy gain and therefore the amount of time spent foraging within a patch is reduced. As a result of the new gain function the average rate of energy gain (R_{y3}) is higher than the previous year.

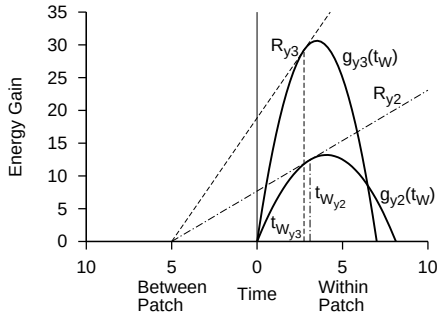


Figure 11.: OFT: Example Year Three Patch

Multiple Patch Types Example

In the previous example the brown bear's environment only consisted of a single patch type. However, as brown bears' forage over large territories spanning thousands of square miles (Garshelis, 2007); it is likely more than one patch type exists in their environment (e.g., forests, rivers). In this example there are two patch types available to the brown bear. Table 7 details the characteristics of each of the patch types. Noticeable is each patch type differs in their time between patches and

gain function.

Table 7: OFT: Example Multiple Patch Types for a Brown Bear

Patch Type	t_B	$g(t_W)$	R	t_W[a]
1	$10\,\text{sec}$	$-0.8 * {t_W}^2 + 6.5 * t_W$	$3.9061\,\text{kCal/sec}$	$1.6212\,\text{sec}$
2	$5\,\text{sec}$	$-2.5 * {t_W}^2 + 17.5 * t_W$	$3.9061\,\text{kCal/sec}$	$2.7188\,\text{sec}$

[a] As calculated from equation 3.6.

When in the specified environment, the bear will spend $1.6212\,\text{seconds}$ in the first patch type and $2.7188\,\text{seconds}$ in the second type. Figure 12 graphically illustrates the optimal time to spend in each patch type and also the average rate of energy gain. As the marginal rates for each patch is the same as the average rate, the tangential lines all have the same slope and are thus parallel to one another.

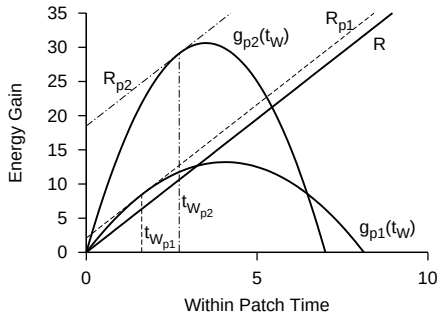


Figure 12.: OFT: Example Optimal Multi-Patch Time

3.2 Adaptive Control of Thought-Rational Theory

The ACT-R theory aims to explain human cognition by (1) describing an architecture of the human mind and (2) the process by which cognition occurs within the stated architecture (Anderson et al., 2004). The theoretical foundation for ACT-R is rational analysis which assumes "...each component of the cognitive system is optimized with respect to demands from the environment, given its computational limitations" (Taatgen and Anderson, 2002, pg. 130). The theory and architecture of ACT-R has been used in research areas such as perception and attention (Byrne, 2001); learning and memory (Fu et al., 2006); problem solving and decision making (Gray et al., 2005);

language processing (Anderson et al., 2001); and other domains relevant to this dissertation, such as information search (Pirolli and Card, 1999).

Figure 13 illustrates the basic architecture of ACT-R 5.0 (Anderson et al., 2004) which consists of modules, buffers, and a central production system. Each module within ACT-R is independent of one another and is responsible for a particular task[11]. The visual and motor modules are part of the perceptual-motor system that interacts with the external environment[12]. The visual module controls vision; attending and identifying objects in the visual space. The manual module directs the hands to perform actions (e.g., picking up an object, clicking a mouse button). The intentional module keeps track of a stack of goals, intentions, and the current state of the problem at hand[13]. Finally, the declarative module interacts with declarative memory (i.e., what is known).

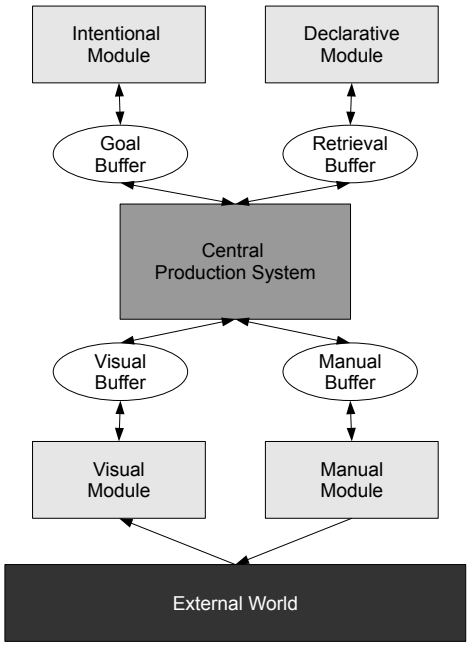


Figure 13.: ACT-R: 5.0 Architecture (Anderson et al., 2004, pg. 1037)

[11]The ACT-R theory does not state the modules listed are the only valid modules used in human cognition (Anderson et al., 2004). Rather, these modules are at the core of the system developed thus far.

[12]Although ACT-R is a foundational theory for IFT, only the higher cognition portion of the theory is used. The perceptual-motor system is not detailed in IFT and thus that portion of ACT-R is only briefly described here. For more information about the perceptual-motor system the reader can refer to Anderson et al. (2004).

[13]A stack is simply a Last In First Out (LIFO) data structure. Whenever a new item is added to the stack it is "pushed" onto the top of the stack. When retrieving an item from a stack the topmost item of the stack is "popped" off the top of the stack.

3.2.1 Central Production System

The central production system (CPS) is responsible for coordinating the activities of all the independent modules (Anderson et al., 2004). The CPS does not directly interact with each module, rather buffers (i.e., working memory) act as intermediaries providing an area for information exchange[14]. Each buffer, however, is limited in capacity to a single chunk of information (i.e., unit of knowledge) (Miller, 1956) at a single time period (Anderson et al., 2004). Such a limitation is to reflect human's limited working memory capacity. For example, only memories being focused on in long-term memory are available at any one time as opposed to having all memories available at all times.

The CPS represents procedural memory (i.e., how to do things) in the form of productions which consist of a set of rules known as production rules. Each production rule consists of a condition or set of conditions and then some action to perform when the condition or conditions are true (i.e., when the conditions match the current state). Figure 14 illustrates six production rules (PR1-PR6) in the form `IF` $< condition(s) >$ `THEN` $< action >$. The CPS uses productions and information from the buffers in order to (1) find rules which match the current state, (2) select the most beneficial rule, and finally (3) execute a rule which results in some action (Anderson et al., 2004).

IF goal is Write-answer
    & answer unknown
THEN set and push subgoal
    Find-solution
    to the goal stack

(a) PR1

IF goal is Write-answer
    & answer unknown
THEN quit and pop
    the goal from
    the goal stack

(b) PR2

IF goal is Write-answer
    & answer known
THEN write answer and pop
    the goal from
    the goal stack

(c) PR3

IF goal is Find-solution
    & answer unknown
    & operation is addition
    & N1 known
    & N2 known
THEN set and push subgoal
    Add-numbers
    to the goal stack

(d) PR4

IF goal is Find-solution
    & answer known
THEN pop the goal
    from the goal stack

(e) PR5

IF goal is Add-numbers
    & N1 known
    & N2 known
THEN retrieve answer and pop
    the goal from
    the goal stack

(f) PR6

Figure 14.: ACT-R: Example Production Rules – adapted from Anderson et al. (2001, pg. 338)

A pattern matching mechanism within the CPS determines if the contents of any of the buffers match the condition of any of the rules. If a match exists the production rule is selected and then

[14]Both the modules and the CPS can read from and write to the corresponding buffer.

fired (i.e., executed). In situations with multiple matching production rules, conflict resolution is undertaken where a conflict set is formed and the rule with the highest probability (based on utility) is selected and executed. The utility of a rule is determined from past experiences of the production rule within the context of the current goal (stored in the goal buffer). The utility of production i is calculated from equation 3.9[15] as U_i (Anderson et al., 2001). P_i is the probability for achieving the goal using production i based on past performance. G is the expected gain from successfully completing the goal independent of the production used. C_i is the average time previous attempts using production i took (i.e., cost) to complete the goal. Finally, ε represents random noise.

$$U_i = P_i G - C_i + \varepsilon \tag{3.9}$$

The actual selection of a production is dependent on its probability as expressed in equation 3.10. U represents the utility of a production and t controls the noise in the utilities (Fu et al., 2006). The actual selection of a production is thus probabilistically-based rather than by absolute utility values.

$$P_i = \frac{e(U_i/t)}{\sum_j^n e(U_j/t)} \tag{3.10}$$

Consider an example of a student attempting to solve the following equation on a math test[16]: $4 + 1$. Using the production rules in figure 14, the cognitive process of the student (broken down at each step in the process) is illustrated in figure 15. The top portion of the figure lists the goals in the goal stack, with the topmost goal signifying the goal in the goal buffer (i.e., the goal currently being attended to). The middle section shows the production rules selected by the CPS to fire. Finally, the bottom portion represents the contents of the retrieval buffer. The values specified for the goal stack and retrieval buffer are representative after the corresponding production has fired.

As the student's overall goal is to write down the answer to the problem the goal *Write-answer* is added to the goal stack. In the first step the CPS performs pattern matching on the current goal and finds two production rules (PR1 and PR2) are valid rules. Since there are two viable candi-

[15] Anderson et al. (2001) does not explicitly include the noise term (i.e., ε) in U_i. However, other ACT-R researchers do (e.g., Fu et al. (2006)) and as the inclusion is more specific the noise term is provided in equation 3.9.

[16] The arithmetic example was adapted from Anderson et al. (2001).

Step	1	2	3	4	5	6
Goal Stack	Write-answer	Find-solution Write-answer	Add-numbers Find-solution Write-answer	Find-solution Write-answer	Write-answer	
Central Production System	Conflict Set {PR1, PR2}	PR1	PR4	PR6	PR5	PR3
Retrieval Buffer				$4 + 1 = 5$		

Figure 15.: ACT-R: Example Cognitive Problem-Solving Process

dates the utility of each production is then calculated. Assuming the utility of PR1 was higher, the student will attempt to solve the problem by adding the goal *Find-solution* to the goal stack in step two. However, part of the process of finding a solution is the summation of the two given numbers (N1 and N2), which is represented in the addition of goal *Add-numbers* at step three. When production PR6 fires in step four, part of the action involves retrieving the chunk representing the addition of 4 and 1 from declarative memory. Once in possession of the answer the goal *Add-numbers* is removed from the stack since it is no longer needed (i.e., the numbers have been added). At step five the current goal is *Find-solution* and since the answer is known production PR5 is fired which removes the goal from the stack. Finally at step six, the only remaining goal is *Write-answer* and since the answer is known production PR3 will fire which will (1) cause the student to write the answer down and (2) remove the goal from the stack, thus ending the cognitive process.

3.2.2 Production Learning

ACT-R is capable of learning new production rules via a mechanism called production compilation (Taatgen and Anderson, 2002). Compilation can occur when two production rules are used in sequence to request and then retrieve a chunk from declarative memory. A single production rule is created which aggregates the two production rules and embeds the declarative knowledge into the rule[17]. Learning in this context removes the potentially expensive operation of chunk retrieval from declarative memory.

To illustrate production compilation, consider the previous example (figure 14) where produc-

[17]When a new production rule is created, the original production rules it was created from are not removed from procedural memory.

tion PR4 requested production PR6 to retrieve an answer from declarative memory. In the example 4 and 1 were provided as numbers to add with the answer of 5 being retrieved from declarative memory. As learning occurs the production rules PR4 and PR6 can be combined for the special case of $4 + 1 = 5$. Therefore, steps three and four from figure 15 are combined, resulting in a reduction of the overall number of steps from six to five. Figure 16 illustrates the new production rule PR7 created through production compilation. Now when $4 + 1$ is encountered the utilities of productions PR4 and PR7 will be compared to determine which production is fired (equations 3.9 and 3.10). The eventual likely outcome is the utility of production PR7 will be higher as the cost does not include (1) the firing of another production and (2) the retrieval of chunk *five* from declarative memory.

IF goal is Find-solution
& answer unknown
& operation is addition
& N1 is 4
& N2 is 1
THEN set answer to 5 and pop
the goal from
the goal stack

(a) PR7

Figure 16.: ACT-R: Example Production Compilation

3.2.3 Chunk

As mentioned in section 3.2.1 a chunk represents a single unit of knowledge (Miller, 1956). The unit of knowledge differs by chunk and can refer to a word, digit, color, shape, phrase, or other such patterns (Simon, 1974). In ACT-R each chunk is of a particular type and associated with slots which represent another chunk or some other value (Stewart and West, 2007). Figure 17 shows an example of three different chunks stored in declarative memory (Anderson et al., 2001; Stewart and West, 2007). Each of the chunks is given a name (e.g., *four-plus-one*, *five*, *large-friendly-dog*) for convenience along with a type (e.g., addition, integer, dog) and some slots. In obtaining the answer to the previous example of $4 + 1$, the chunk *four-plus-one* would have been activated which would have lead to the retrieval of chunk *five* (since the sum slot of *four-plus-one* refers to the *five* chunk). If a person was trying to recall knowledge about a large, friendly dog instead, chunk *large-friendly-dog* would be retrieved.

Chunk: four-plus-one

isa	addition
addend1	four
addend2	one
sum	five

(a) Four-plus-one

Chunk: five

isa	integer
value	5

(b) Five

Chunk: large-friendly-dog

isa	dog
size	large
manner	friendly

(c) Large-friendly-dog

Figure 17.: ACT-R: Example Chunks (Anderson et al., 2001; Stewart and West, 2007)

3.2.4 Declarative Memory

As seen in step four of figure 15, retrieving information from long-term memory is an important process of human cognition. Within ACT-R declarative knowledge is encoded as a network structure (Anderson and Pirolli, 1984). The network consists of nodes (i.e., chunks) connected via links (Collins and Loftus, 1975). Links are determined based on the association between nodes. Strongly associated nodes are located in close proximity to one another, whereas weakly associated nodes are distal from one another. Nodes may also be indirectly associated via intermediary nodes. Figure 18 provides an example of the network structure found in declarative memory. Each ellipse represents a node, while each line is a link and thus represents an association between nodes.

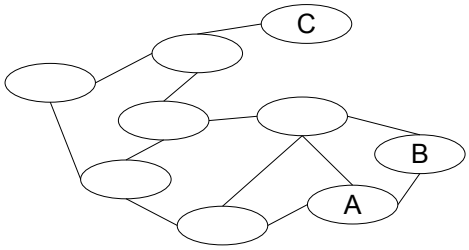


Figure 18.: ACT-R: Declarative Memory Network Structure

Spreading activation is the process by which chunks related to a given source chunk can be retrieved from memory (Collins and Loftus, 1975). When some cue is attended to (e.g., when a user reads a particular word) the chunk j representing that cue is activated in memory (Anderson and Pirolli, 1984). The activation then spreads from the source of the activation (i.e., the cue) throughout the entire network activating any associated nodes. The spreading occurs instantaneously throughout the network and the strength of activation at each node decays exponentially with distance from the source (Anderson and Pirolli, 1984). The end result is more strongly acti-

vated nodes represent knowledge which is more relevant to the activation source.

The total activation associated with chunk i (i.e., a chunk in declarative memory) when chunk j is the source activation is expressed in equation 3.11[18] as A_i (Anderson et al., 2004). B_i is the base-level activation which takes into account the history of chunk i independent of chunk j (Anderson and Milson, 1989). The base-level activation is dependent on the frequency and recency of prior activations of chunk i (Anderson et al., 2004) and follows a power law of learning and forgetting where activation strength increases with recent and repeated usage (Anderson et al., 2001). W_j is a weight representing the amount of attention being paid to the source chunk j. S_{ji} is the strength of association between chunks j and i. Finally, ε is noise associated with the activation process.

$$A_i = B_i + \sum_j W_j S_{ji} + \varepsilon \tag{3.11}$$

The manner in which spreading activation applies to information retrieval is based on the strength of activation for the chunk to be retrieved (chunk i) when the source of activation is the proximal cue chunk j. The strength of a chunk's activation determines "...its probability of being retrieved and its speed of retrieval" (Anderson et al., 2004, pg. 1042). Therefore, if chunk i has weak activation it may (1) not be retrieved or (2) take too long to retrieve. However, absolute activation strength does not guarantee chunk retrieval since each chunk has a retrieval probability as expressed in equation 3.12 (Anderson et al., 2004). In equation 3.12, A_i is the total activation of chunk i, τ represents a threshold which the activation must be above, and s is related to the variance of activation noise (Anderson et al., 2004).

$$P_i = \frac{1}{1 + e^{-(A_i - \tau)/s}} \tag{3.12}$$

Assume in figure 18 that nodes A, B, and C represent chunks *four-plus-one*, *five*, and *large-friendly-dog*, respectively from figure 17. At step four of the student's process for solving the equation $4+1$ (figure 15), the source of activation would have been chunk *four-plus-one*. Based on the given network structure the chunks *four-plus-one* and *five* are directly and closely associated with one another indicating some degree of similarity. Therefore, the total activation of chunk *five*

[18] Anderson et al. (2004) does not explicitly include the noise term (i.e., ε) in A_i. However, other ACT-R researchers do (e.g., Fu et al. (2006)) and as the inclusion is more specific the noise term is provided in equation 3.11.

would likely be high and probably lead to a successful retrieval of the chunk. The same success would be less likely if the *large-friendly-dog* chunk was to be retrieved given chunk *four-plus-one* as the source of activation. A weaker activation would be likely since the distance between the chunks is greater and there are no direct associations[19].

3.3 Information Foraging Theory

The theory of information foraging is concerned with not only the way a person searches within their environment, but also how the environment can be shaped to better facilitate foraging. Therefore, research has used IFT to not only look at navigational patterns of foragers, but also how information environments can be altered to facilitate foraging. IFT has been used to inform the design of graphical user interface controls (e.g., checkboxes, list boxes) which provide social activity visualizations as navigational cues (Willett et al., 2007); highlight ScentTrails on Web pages which facilitate a user's search for information (Olston and Chi, 2003); find optimal browsing paths for large pictures displayed in limited viewing areas (Xie et al., 2006); explain navigational choices within source code during program maintenance tasks (Lawrance et al., 2007); describe the effects of delay, familiarity, and breadth on users' performance, attitude, and intentions at Web sites (Galletta et al., 2006); and the role of scent in the decision to browse a menu as opposed to searching a Web site (Katz and Byrne, 2003).

The foundational theories OFT and ACT-R are used by IFT to explain the behavioral and cognitive aspects of information foraging. Like OFT, the same sequential search–encounter–decision process is used to explain the basic behaviors of an information forager. Similar to how animals search for patches using their sense of smell, information foragers use a metaphorical sense of smell to locate and follow an information scent trail. The mechanism by which this information scent works is explained via the ACT-R theory. Once an information patch (i.e., an item of interest) has been located the decision turns to answering the question of "how long to stay in a patch?" from the classical patch model. ACT-R also explains the details of how the decision of when to

[19] Although the link between an addition problem and a large, friendly dog seems totally unrelated, such associations may exist within a person's mind. Thus the activation chunk *four-plus-one* may in fact allow retrieval of chunk *large-friendly-dog* especially when taking into account the probabilistic nature of chunk retrieval. For example, the summation problem may lead to an association with summer. Summer may be associated with summer breaks from school which in turn is associated with early childhood. Childhood may then be associated with the family pet that was in turn a large, friendly dog.

stay or leave a patch is determined by the information forager.

The following sections present an in-depth explanation of the concepts of information scent and information patches followed by a description of two versions of an IFT model (SNIF-ACT 1.0 and 2.0).

3.3.1 Information Scent

Information scent is "the detection and use of cues, such as World Wide Web (Web) links . . . that provide users with concise information about content that is not immediately available" (Pirolli, 2007, pg. 68). The concept of information scent corresponds to the search portion of the search–encounter–decision process from OFT. Just as in the wild, a lack of scent makes the probability of encountering the item of interest difficult. However, unlike animals hunting for food where one berry is just as beneficial as another berry, information is not as interchangeable. Rather, information of value should be (1) relevant to an information forager's goal and (2) novel (Sandstrom, 1994).

The two main ways in which information scent is used is to (1) guide users to the information being sought and (2) provide a general impression of the available content within a patch. In a Web environment cues are obtained from the text and images associated with a hyperlink. The predicted utility of a link is based on how the cues from a link match a user's goal (i.e., the probability of a link providing a Web page with the desired information[20]). The link with the highest predicted utility (i.e., scent) is then selected as the next navigational choice.

The scent of a link is based on the goal of a user. The user's goal G is the desired distal information where i represents each goal feature (i.e., each word of a goal). Each proximal cue (i.e., link), L, on the Web page indicates the distal content of the linked page, where j represents each cue feature (i.e., each word of the link)[21]. The features for both G and L are represented cognitively as chunks (Miller, 1956).

The value of link L in the context of goal G is expressed in equation 3.13 as the sum activation (equation 3.11) of each goal feature (Pirolli, 2007).

[20] Such a relationship between link text and the content of the linked page has been demonstrated empirically by Davison (2000).

[21] Common stop words from hyperlinks such as *and*, *the*, *a*, etc. are not included as features of a cue.

$$V_{L|G} = \sum_{i \in G} A_i \tag{3.13}$$

The choice of which link to select is based on the link with the highest utility within the context of the goal G. Expressed in equation 3.14, the utility is the value of link L (equation 3.13) along with a random component ε (Pirolli, 2007). The random component represents user and context variability.

$$U_{L|G} = V_{L|G} + \varepsilon_{L|G} \tag{3.14}$$

Similar in the way that ACT-R selects which production to fire probabilistically, IFT determines which link to select via probability. Equation 3.15 is the probability of selecting a given link L from a set of links C within the context of goal G (Pirolli, 2007). $U_{L|G}$ is the link being evaluated, $U_{k|G}$ represents the utility for each link in the set, and μ reflects a scaling parameter for random noise.

$$Pr(L|C,G) = \frac{e^{\frac{U_{L|G}}{\mu}}}{\sum_{k \in C} e^{\frac{U_{k|G}}{\mu}}} \tag{3.15}$$

To illustrate the concept of information scent, consider the following example of a person searching for information on the Web. The goal (G) of the user is to find information regarding "white lily flowers." Figure 19 represents the relevant fragment of the user's declarative memory where each feature chunk i of the goal is represented as a black ellipse. On the current Web page the user is presented with two links "red roses" (L_1) and "cherry trees" (L_2). The chunk features of links L_1 and L_2 are symbolized as light gray and dark gray ellipses in figure 19, respectively. As seen in figure 19 the features of L_1 are closer than L_2 to the goal chunks and thus more similar and more likely to strongly activate the features of the goal. Therefore, the scent of link L_1 is stronger (i.e., has a higher utility) and the user will select the first link[22].

Although based on ACT-R, the concept of information scent in IFT deviates from the ACT-R theory in three main ways (Pirolli, 2007). First, the source of activation in ACT-R is the goal chunk. In IFT the chunk representing the feature of a proximal cue is the source and the goal

[22]This example assumes the noise from equation 3.11 and the random component of equation 3.14 are comparable across link features and links.

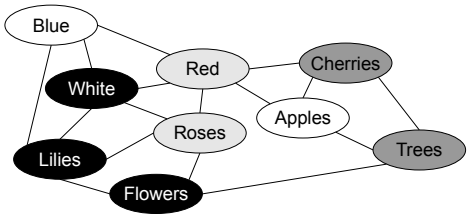


Figure 19.: ACT-R: Example Memory Schematic – adapted from Collins and Loftus (1975, pg. 412)

chunk is the destination. Second, the purpose of spreading activation in ACT-R is to retrieve some chunk from declarative memory. IFT is not interested in the retrieval of a chunk, but rather the total level of activation on a goal chunk. Lastly, the utility of which link to select is not based on past successes and failures like productions in ACT-R are. Instead, utility is based on total activation strength of a link which does not take into account past performance. A lack of history therefore means knowledge of previously successful associations between links and success are not considered. For example, the utility of the link "contact us" would be made independent of any prior successes a user has had when clicking on a similarly named link to find contact information for a Web site.

3.3.2 Information Patch

An information patch is a grouping of information where "... it is easier to navigate and process information that resides within the same patch than to navigate and process information across patches" (Pirolli, 2007, pg. 49). Within a Web context what constitutes an information patch can differ depending on the level of analysis. At a high-level an individual Web site could be considered a patch whereas at a lower-level the Web pages within a single Web site could each be considered a patch. Prior research has not explicitly made distinctions between patches at differing levels of analysis. In order to be clear, the terms *site-patch* and *page-patch* will refer to patches which constitute an entire Web site or Web page within a site, respectively.

Although the definition of what a patch is differs by level of analysis, the relationship of similarity within and across patches does not differ. For example, information within a Web page is more similar than across Web pages of the same site which in turn, is more similar than Web pages of another site. Likewise, a single Web site will have more coinciding information compared to

another Web site.

Such a topical patchy structure of the Web has been empirically demonstrated (Davison, 2000) strengthening the logical argument for a patchy Web. The similarity of content between pages from the most similar to least were (1) linked pages within the same domain; (2) unlinked pages within the same domain; (3) linked pages to different domains; and finally (4) random pages[23] (Davison, 2000). An increase in link distance (i.e., degrees of separation) within the same domain has also been associated with decreases in page content similarity (Pirolli, 2007). The aforementioned research lends support to the assertion of a patchy grouping of information on the Internet where patches in "close proximity" to one another are more similar than patches farther apart.

As the Internet exhibits a patchy structure, the patch model from OFT (Charnov, 1976) is appropriate to use and can determine the optimal length of stay for a forager. The decision to stay in or leave a patch is determined such that a person will "... forage in an information patch until the expected potential of that patch is less than the mean expected value of going to a new patch" (Pirolli, 2007, pg. 81). The patch leaving rule is mathematically stated in equation 3.16 (Pirolli, 2007) as a variant of the Marginal Value Theorem (Charnov, 1976). $U(x)$ is the utility of a forager in their current state and $\bar{U}$ is the mean utility of other patches. Thus just like the marginal value theorem, the visitor will continue to forage in the patch as long as the utility (i.e., marginal value) is higher than the average of all other patches (i.e., average rate of return). Once the current state utility is equal to the mean, the forager will leave the patch.

$$U(x) > \bar{U} \tag{3.16}$$

3.3.3 SNIF-ACT

Pirolli (2007) implemented two versions of a model based on IFT called SNIF-ACT (Scent-based Navigation and Information Foraging in the ACT architecture). As the ACT architecture is a major component of IFT, a set of production rules were defined for both versions of the SNIF-ACT models which characterized users' actions while foraging. Figure 20 lists each of the pertinent productions showing how a user starts processing a new page; evaluates links on a page; and decides amongst clicking a link, going back to a previous page, or leaving the site.

[23] Similarity within a single page is not included since by definition no page can be more similar to a page than itself.

The first four productions of figure 20 (*Start-process-page*, *Process-links-on-a-page*, *Attend-to-link*, *Read-and-evaluate-link*) are concerned with the cognitive aspects of reading, attending, and evaluating links on a page. In terms of OFT, if any of the first four productions fire then that is a decision by the visitor to continue foraging within the same page-patch. The final three productions also relate to the patch-leaving rule of OFT, although their levels of analysis differ. Production *Click-link* represents either a decision to leave a page-patch or site-patch depending on if the link was internal to the Web site or external. Production *Leave-site* relates to the leaving of a site-patch and production *Backup-a-page* is concerned with going to an already visited page-patch. In any case, a production representing the leaving of either a page-patch or site-patch should fire when the marginal value of the patch drops to the average rate of return for all patches.

IF goal is Start-next-patch
& there is a task description
& there is a browser
& browser on unprocessed page
THEN set and push subgoal
Process-page to the goal stack

(a) Start-process-page

IF goal is Process-page
& there is a task description
& there is a browser
& there is an unprocessed link
THEN set and push subgoal
Process-link to the goal stack

(b) Process-links-on-page

IF goal is Process-link
& there is a task description
& there is a browser
& there is an unattended link
THEN choose an unattended link and
attend to it

(c) Attend-to-link

IF goal is Process-link
& there is a task description
& there is a browser
& the current attention is on a link
THEN read and evaluate the link

(d) Read-and-evaluate-link

IF goal is Process-link
& there is a task description
& there is a browser
& there is an evaluated link
& the link has highest activation
THEN click on the link

(e) Click-link

IF goal is Process-link
& there is a task description
& there is a browser
& there is an evaluated link
& the mean activation on page is low
THEN leave the site and
pop the goal from the goal stack

(f) Leave-site

IF goal is Process-link
& there is a task description
& there is a browser
& there is an evaluated link
& the mean activation on page is low
THEN go back to the previous page

(g) Backup-a-page

Figure 20.: SNIF-ACT: Production Rules (Pirolli, 2007, pg. 97)

Also fundamental to IFT is the concept of information scent which is determined by the level of activation (equation 3.11) of a goal from a given link. The base-level activation of the goal chunk (B_i) in both versions of SNIF-ACT was assumed to be static (i.e., the goal did not change) and thus B_i was set to zero (Pirolli, 2007). The amount of attention paid to a link cue (W_j) was modeled as exponentially decaying with respect to the number of cues in a link as shown in equation 3.17 (Pirolli, 2007). W and d are scaling factors and n is the number of cues (i.e., words) in a link.

$$W_j = We^{-dn} \tag{3.17}$$

To determine similarity between a cue and goal feature (S_{ji}), a measure from information theory known as pointwise mutual information (PMI) was used (Church and Hanks, 1989). The formula for PMI (equation 3.18) determines the association between two words i and j (or in IFT between a cue and goal feature). The numerator in equation 3.18 is the probability of the two words occurring together whereas the denominator specifies the probability of the words occurring independently. When normalized, a PMI score of 0 indicates no association whereas as a score of 1 means perfect association between the two words. PMI has been found to be a good proximal measure of the associations a person may make between chunks within their own declarative memory. For example, PMI was more accurate on tests of synonymy than typical college applicants taking the Test of English as a Foreign Language (TOEFL) (Turney, 2001).

$$PMI(i,j) = log\left[\frac{Pr(ij)}{Pr(i)Pr(j)}\right] \tag{3.18}$$

SNIF-ACT 1.0

The first version of SNIF-ACT assumed foragers evaluated all links on a page before deciding which link to select (Pirolli, 2007). The model was tested against protocol data collected from Card et al. (2001). Four student subjects were given two experimental information finding Web tasks. The first task required the subject to obtain the date and a picture of a comedy group performing at a college campus. For the second task, subjects were instructed to find four posters from the movie Antz. The keystrokes, mouse movements, eye movements, Web pages visited, and think-aloud comments were captured from each subject.

The model was evaluated on the ability of information scent to predict link-following and site-leaving actions. From the eight datasets (four subjects with two tasks a piece), a total of 91 link clicks were captured. Using the concept of information scent, the SNIF-ACT model's prediction of which link would be followed was found to be significantly different from random selection ($\chi^2(30) = 18{,}589.45$; $p < 0.0001$) (Pirolli, 2007). Such a result lends credence to the idea of information scent being an indicator used by people to locate proximal information.

In terms of site-leaving, the SNIF-ACT model was also found to follow the patch-leaving rule whereas the subjects foraged in a site-patch until the "...expected potential of that patch is less than the mean expected value of going to a new patch" (Pirolli, 2007, pg. 81). Figure 21 illustrates how a drop in information scent can be a cue for the value of a site-patch. The scent of the last page visited at a Web site was, on average, lower than the average scent of the first page of a new Web site. Therefore, this lack of scent was an indicator to the subject that this site-patch does not contain the sought after goal information.

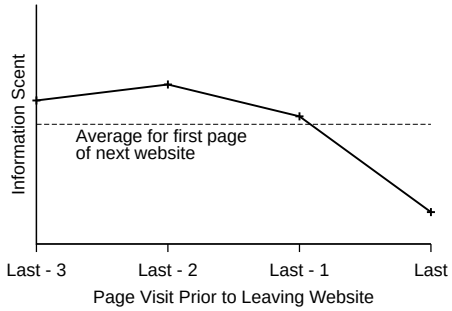


Figure 21.: SNIF-ACT: Site-leaving Actions (Pirolli, 2007, pg. 100)

SNIF-ACT 2.0

The second version of SNIF-ACT removed the unrealistic assumption from SNIF-ACT 1.0 that foragers would attend to and evaluate each link before making a decision of where to go. Instead, a learning mechanism was used which relied on the concept of satisficing (Simon, 1956). As a forager has imperfect information and limited computational facilities an optimal decision is unlikely. However, a decision which satisfies a need at some specified level is probable. Therefore, with regards to satisficing the forager would continue to evaluate links in SNIF-ACT until a "good enough" link was found (even though the link might not be optimal). The determination of what is

"good enough" relies on the ability of the forager to learn as information is uncovered while foraging.

In order to implement such a learning mechanism the utilities and probabilities of productions *Attend-to-link*, *Click-link*, and *Backup-a-page* were updated to include the history of links already attended to and pages already visited[24] (Pirolli, 2007). After evaluating a link, the forager is faced with the decision of whether to attend to the next link, click a previously evaluated link, or leave the page. Determined from each production's utility (equations 3.19-3.21), the forager's ultimate decision is based on the probabilities for each production (equations 3.22-3.24) (Pirolli, 2007).

Equation 3.19 is the utility for production *Attend-to-link* where $U_{L|G}$ represents the utility of the current link (equation 3.14) and n is the current number of links already evaluated.

$$U_A(n+1) = \frac{U_A(n) + U_{L|G}}{1+n} \tag{3.19}$$

The utility for production *Click-link* is shown in equation 3.20 where $max(U_{L|G})$ is the maximum link utility of the links evaluated so far and k is a scaling parameter.

$$U_C(n+1) = \frac{U_C(n) + max(U_{L|G})}{1+k+n} \tag{3.20}$$

Taking into account the value of prior pages and the cost of backing up, equation 3.21 represents the utility for production *Backup-a-page*. $\bar{U}_{Page}$ is the average utility of previously visited pages (within the same Web site), $\bar{U}_L(n)$ is the average link utility of links 1 to n, and C_{Back} reflects the cost of returning to a previous page.

$$U_B(n+1) = \bar{U}_{Page} - \bar{U}_L(n) - C_{Back} \tag{3.21}$$

The probabilities for each of the three production rules are expressed in equations 3.22-3.24. Each equation is the probability of selecting the given production after the evaluation of n links on a page. In each equation, μ represents a scaling parameter.

$$Pr(\textit{Attend-to-link}, n) = \frac{\exp\left[\frac{U_A(n)}{\mu}\right]}{\exp\left[\frac{U_A(n)}{\mu}\right] + \exp\left[\frac{U_B(n)}{\mu}\right] + \exp\left[\frac{U_C(n)}{\mu}\right]} \tag{3.22}$$

[24]Production *Leave-site* was not updated since the experiment using SNIF-ACT 2.0 took place on a single Web site.

$$Pr(\textit{Click-link}, n) = \frac{\exp\left[\frac{U_C(n)}{\mu}\right]}{\exp\left[\frac{U_A(n)}{\mu}\right] + \exp\left[\frac{U_B(n)}{\mu}\right] + \exp\left[\frac{U_C(n)}{\mu}\right]} \tag{3.23}$$

$$Pr(\textit{Backup-a-page}, n) = \frac{\exp\left[\frac{U_B(n)}{\mu}\right]}{\exp\left[\frac{U_A(n)}{\mu}\right] + \exp\left[\frac{U_B(n)}{\mu}\right] + \exp\left[\frac{U_C(n)}{\mu}\right]} \tag{3.24}$$

Example

To better visualize the interplay amongst the three production rules, a hypothetical example of a Web page with 15 links is provided. The distribution of link utilities ($U_{L|G}$) is defined by the function $15e^{-0.7x} + 1$ and shown graphically in figure 22. Noticeable is the sharp decline in scent from the first link ($U_{L_1|G} = 8.4488$) to the last link ($U_{L_{15}|G} = 1.0004$) on the page.

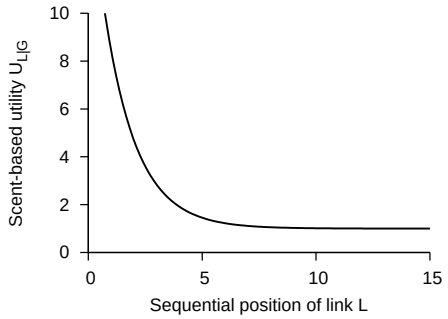


Figure 22.: SNIF-ACT: Hypothetical Distribution of Link Utilities (Pirolli, 2007)

To simulate the utilities and probabilities for each production the following measures were set in accordance with Pirolli (2007): k was set to 5 (equation 3.20); $\bar{U}_{Page}$ and C_{Back} were set to 10 and 5 (equation 3.21); and μ was set to 1 (equations 3.22-3.24). The probability of a forager choosing from each of the three productions given the stated link utility distribution is illustrated in figure 23.

In figure 23 the probability of attending to the next link is high when only a couple links have already been evaluated. This represents the forager learning the value of the current page's links. After more links are evaluated ($n \approx 4$) the forager is better informed of the existence of any highly-scented links which may lead to a goal. Therefore, the probability of clicking on a link rises to its highest level. However, each successive link's scent ($n \gtrsim 5$) drops and begins leveling off near the minimum scent value. Considering none of the previous links were satisfactory

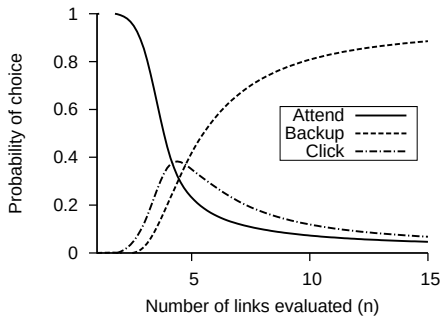


Figure 23.: SNIF-ACT: Hypothetical Production Probabilities (Pirolli, 2007)

in causing the forager to click on them, the likelihood of any of the remaining low-scent links causing a link-following action is low (as evidenced by the decline in the probability of clicking a link). As the forager reaches the end of the links, the probability of the visitor returning to a previously visited page continues to increase.

Model Validation

The SNIF-ACT 2.0 model was tested against data from Chi et al. (2003) for fit to both link-following actions and the decision of foragers to go back a page. 244 subjects were recruited to complete some portion of a total of 32 information foraging tasks on four different Web sites (eight tasks per site). To test the SNIF-ACT 2.0 model, Pirolli (2007) included 74 subjects who completed the tasks at two of the Web sites. The Web sites were chosen due to the static nature of their Web pages (i.e., the content and links of the pages did not change dynamically). Eight of the tasks took place on Yahoo!'s help Web site while the remaining eight occurred on ParcWeb's internal company intranet. Unlike SNIF-ACT 1.0, which was tested against individual clickstreams, the aggregated statistics of the subjects and the SNIF-ACT 2.0 model were compared.

Using linear regression the fit of the aggregated SNIF-ACT model obtained good fit for both the ParcWeb tasks ($R^2 = 0.72$) and the Yahoo! tasks ($R^2 = 0.90$) (Pirolli, 2007). The high R^2 further bolsters the support found in SNIF-ACT 1.0 that information scent is a reliable indicator of the navigational choices a visitor makes when foraging. To test for subjects returning to a previous page another linear regression model was created. Similar to the link-following results, good fit was also found for the ParcWeb tasks ($R^2 = 0.73$) and the Yahoo! tasks ($R^2 = 0.80$) (Pirolli, 2007). Since backing up a page is concerned with leaving a patch at the page-patch level, the re-

sults do not directly bolster the results found in SNIF-ACT 1.0 (which looked at the site-patch level). Instead, the results provide initial supporting evidence of the patch-leaving rule at the page-patch level.

3.4 Conclusion

The preceding sections provided a thorough review of OFT, ACT-R, and IFT. Since IFT draws quite heavily from both OFT and ACT-R, details of each theory was included to give a more complete understanding of IFT. Specifically, the basic prey and patch models from OFT and the architecture and mechanisms for cognition from ACT-R were described. A discussion of information scent and patches, in regards to IFT, was then given along with the details of the SNIF-ACT 1.0 and 2.0 models.

Chapter 4

Hypotheses

Information foraging theory (IFT) is concerned with the information gathering *search* process. However, investigating goal achievement on long tail Web sites is focused on how information gathering characteristics can be used to predict action, such as submitting a contact form. In order for the possibility of action to occur, a visitor must move beyond the information gathering search stage to a decision-making point where an action may or may not take place. Therefore, the information gathering characteristics which are likely to lead to a conversion are those which bring a visitor closer to meeting their information requirements.

Figure 24 illustrates how IFT is used within the decision making process.

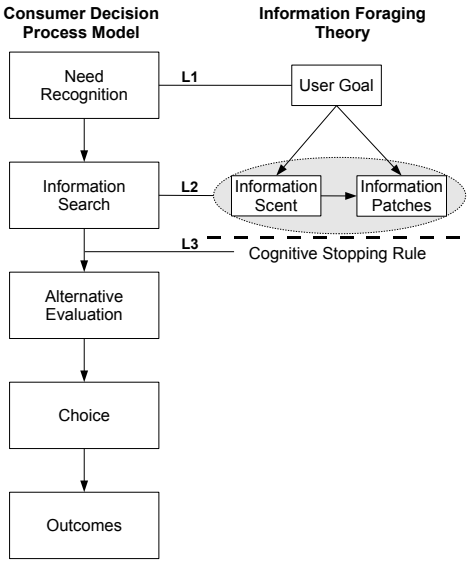


Figure 24.: Consumer Decision Process Model and Information Foraging Theory

On the left-hand side of figure 24 is the consumer decision-making process (CDP) model (Engel et al., 1990). The purpose of the CDP model is to illustrate the basic stages a consumer goes

through when faced with a decision. Although the stages are depicted linearly in the figure, the process itself may be an iterative one.

The decision making process begins when a consumer recognizes some need to be met. In order to fulfill their need, the consumer will then search for information to find possible solutions. In the next stage, each of the potential alternatives found are evaluated against one another until a single alternative is selected. In the final stage, the consumer reflects on the outcomes of the process.

On the right-hand side of figure 24 are the main concepts of IFT. In IFT, user goals initiate and drive the search process. Thus, the goal of the user affects the scent of every cue encountered and how a patch is judged. In addition, the scent of a link also affects which patches will be selected to forage within.

The manner in which IFT applies to the CDP model is shown via lines L1 and L2 in figure 24. Line L1 demonstrates that the need being recognized in CDP is the user goal in IFT. This need or goal is the reason for the process or foraging to occur. In addition, the need or goal sets the context for all subsequent activity. Line L2 illustrates that information scent and patches are concerned with the information search process. Information scent is used to locate patches of information and foraging within a patch obtains relevant information.

Noticeable is how IFT only applies to the first two stages of the CPT model. However, the possibility of a goal being achieved on a Web site can only occur in the fourth stage, once a choice has been made. In order to get to the fourth stage, enough information must first be gathered and any alternatives need to be evaluated. The termination of the information search process occurs at "... some point because the person judges that he has enough information to move to the next stage in the problem-solving or decision-making process" (Browne et al., 2007, pg. 91).

The determination of when enough information has been gathered is via a cognitive stopping rule (Browne et al., 2007) as illustrated by line L3 in figure 24. The cognitive stopping rule may be concerned with the fulfillment of a single criterion, list of items, amount of information, amount of new information, or when understanding of the information stabilizes (Browne et al., 2007; Pitts and Browne, 2004). Regardless of the cognitive stopping rule a visitor uses to judge the sufficiency of their gathered information, some rule must be met before there is a chance of a goal occurring.

Once a forager has stopped collecting information, the alternatives are evaluated. The alterna-

tives may be between multiple products or services; or simply between selecting this product or service or not. If the choice is made for some product or service then the forager may perform some action (e.g., submitting contact information); if not, the forager may leave the site looking for more information and other alternatives.

The remainder of this chapter is outlined as follows: first a brief review is given of the way in which an information forager browses. Next, the user- and site-centric clickstream models of information foraging (CMIF) are introduced. The hypotheses generated from the models to help answer research question 3 (listed below) are then presented in §4.2.1 for the user-centric model and §4.2.2 for the site-centric model.

Research Question 3: *How can information foraging theory and clickstream data be used to explain the achievement of a goal at a long tail Web site?*

4.1 Information Foraging

The basic way in which an information forager evaluates every Web page they are presented with is explained in the first subsection below. An example session is then shown of a user's clickstream as they hunt for information over multiple Web sites. Using the concepts of information foraging, the rationale behind the user's browsing behavior is provided.

4.1.1 Page Evaluation

When presented with a page, a forager has four basic actions which can be selected at any particular time: (1) evaluate or continue to evaluate the links on a page; (2) click on an already evaluated link; (3) go to a previously visited page; or (4) leave the site (Pirolli, 2007). The probability of what action a forager will choose changes over time. When first presented with a new page, it is more probable that the user will begin evaluating the page compared to the other three actions. The purpose of evaluation is to get a general sense of the value of the page and its links.

With continued evaluation, it is likely the probability of at least one of the other three actions becomes higher than the probability for further evaluation. This change in probabilities is due to the concept of satisficing (Simon, 1956; Pirolli, 2007), where the user will continue to evaluate a page until a link with a "good enough" scent is found or it is determined the page does not contain

any “good enough” links. If a highly-scented link is found, it will be clicked. If not, the user will either backup to a previously visited page or leave the site in search of a Web site with a higher mean expected value than the current site (Pirolli, 2007).

The rationale behind why scent is beneficial to a user is due to the costs associated with browsing. Foragers are assumed to be rational and thus try to reduce their search costs while hunting for information (Pirolli, 2007). As each additional page viewed incurs a search cost, taking meandering, wrong, or already traversed paths is less efficient than taking a direct path to the information sought. Information scent is a mechanism by which foragers are able to reduce their search costs by increasing their accuracy on which option leads to information of value. Therefore, a forager will click a link if the scent is deemed high enough to efficiently lead to valuable information. If none of the links provide sufficiently high scent, the forager will perform one of the other three actions in anticipation the action will lead to higher-scented links.

4.1.2 Sample Session

By definition, long tail Web sites do not generate heavy traffic. Their relative obscurity means it is unlikely many new visitors will know of the site’s existence let alone its uniform resource locator (URL). However, the widespread use of search engines by Internet users (comScore, Inc., 2007a) provides a gateway to these long tail Web sites. The results from search engines also provide links to a number of other known and unknown Web sites too. Therefore, an information forager has easy access to numerous Web sites when hunting for information.

Figure 25 shows an example of the clickstream of a successful foraging trip by a user searching for information about an upcoming gig for a comedy troupe at a college campus (Card et al., 2001). The figure is an adaptation of a Web behavior graph (WBG) (Card et al., 2001) which illustrates each Web page visited by a user. The figure is meant to be read left to right and top to bottom. Each rectangular box represents a Web page and each rounded box represents the results returned from a search query. The letter in each box is the Web site and the number is the Web page at that site. All the boxes from the same Web site are shaded the same color. Straight arrows represent the user clicking a link from one Web page to another. Curved arrows at the end of a line represent a user returning to whatever Web page is listed first on the next row down. Vertical lines indicate a return to a previously visited Web page. Figure 25 is a graphical representation of the

following clickstream:

$< A1, A2, A1, B1, B2, C1, D1, D2, D3, D1, D3, D2, D1, C1, D1, C1, B2, E1, E2, E3, E4 >.$

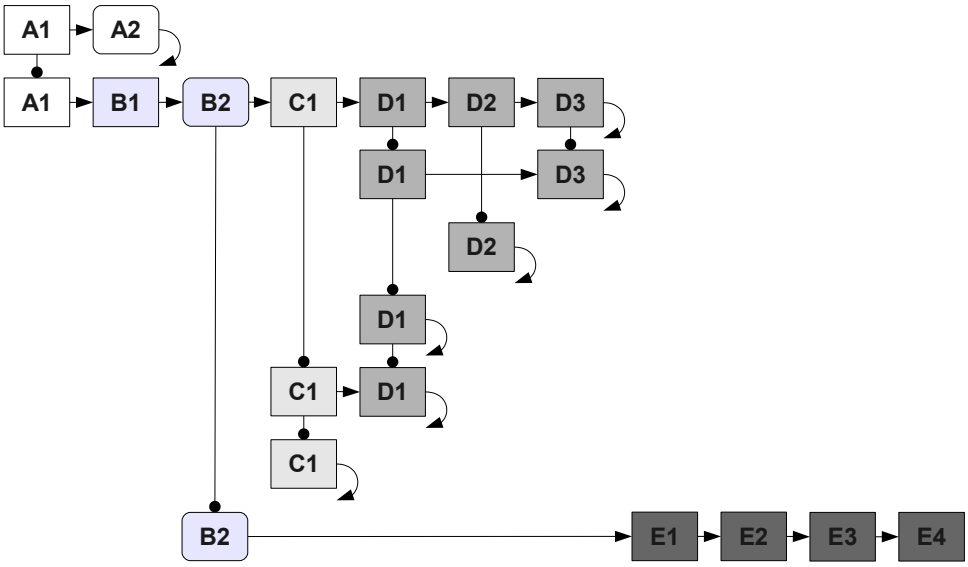


Figure 25.: User-centric: Example User Clickstream Graph – Adapted from (Card et al., 2001)

The clickstream illustrated in figure 25 is user-centric in nature since it includes the browsing behavior of the forager across every Web site the user visited (Padmanabhan et al., 2001). A site-centric version of this clickstream would only include the browsing behavior at a single Web site without knowledge of what occurred at the other Web sites. The term *user-session* refers to a time-contiguous sequence of Web pages viewed at any Web site from the same user, such as seen in a user-centric clickstream. In contrast, *session* represents a time-contiguous sequence of Web pages viewed at the *same* Web site by the same user, like in a site-centric clickstream.

Foraging Explanation

In figure 25 the user started their user-session within patch A (i.e., Web site A) at page A1, a Web page with search capabilities, and entered a search query. After evaluating the results of the query on A2, none of the resulting links had a high enough scent to warrant clicking on and thus the user returned to the first page. Re-evaluating the value of the patch in light of page A1 and the results returned on A2, the user decided to leave the site for patch B.

Site B also had search capabilities and the user again entered a search query. This time, while evaluating the results of the query on B2, one of the links had a high enough scent to cause the

user to click on it. On site C the user found a highly-scented link to site D and clicked that link. On site D, the user can be seen as having relatively poor scent due to the inefficient revisiting of a number of pages (D1, D2, and D3) multiple times.

After determining that the value of patch D had dropped below what could be expected elsewhere, the user returned back to the previous patch (site C) and finally back to the results page of site B. Re-evaluating the links on the results page B2 lead the user to select another link which lead to site E. The scent throughout site E was strong as the user did not backtrack. In addition, the user found the information they sought about the comedy troupe on page E4.

The preceding example illustrated how concepts from IFT can be used to explain users' browsing behaviors. For example, a lack of highly scented links from any of the results returned on page A2 explains why the user backtracked to page A1 after executing a search. In addition, the movement from Web site A to B can be deciphered as the user believing information of greater value could be obtained from another patch. The next section presents a clickstream model developed from IFT which captures these concepts using clickstream metrics.

4.2 Clickstream Model of Information Foraging

The clickstream model of information foraging uses clickstream metrics to represent the concepts of information scent and patches. The user-centric (UC) model is presented first which uses information about a forager's entire browsing behavior to determine the overall scent at and the value of a Web site. Since data about a user's entire clickstream is rarely available, a site-centric (SC) version of the model is also presented which provides alternative conceptualization of the IFT concepts using only site-centric data.

4.2.1 User-centric

Of the four possible actions a user may take at any point on any page, only three of those actions are directly observable via a user's clickstream: click on a link, return to a previously visited page, and leave the site. Although the determination of scent is internally represented as the activation between the features of the links and goal (Pirolli, 2007), the observable actions of a user's clickstream can be used as proxies for determining how a user perceived scent and judged the value of a patch.

The judgment of information scent or the value of the patch cannot, however, be determined in absolute terms from a user's clickstream since there is no absolute to compare against. Rather, judgment must be done in relative terms. For example, assume a user is on a page which has one link that goes to page A and another to page B. If the user's clickstream shows page A was visited next then the link to page A had higher scent than the link to page B. The actual scent and thus the difference in scent between the links are unknown.

Potentially more important than the scent of each individual link, however, is the overall scent and patch value at a particular Web site in comparison to the other Web sites visited. Such a means of determining which Web site was of more value to a user provides a clue into which site might fulfill the goal of the user. For example, in figure 25 the user visited three pages multiple times on site C which would indicate a poor scent at the site. Contrast that browsing behavior with site E where four pages were visited only a single time.

The relative judgment between sites is also important in cases where the user's information goal is complex. For such goals it is likely the clickstream of a user will be complex regardless of the site being visited. If judged in absolute terms, it would seem unlikely the user would find the information they sought at any site. However, if judged relatively, it may be found that one site, while still having an overall low scent, has a higher scent than the other sites and thus was the most useful.

For example, assume a user visited 15 pages at one site, with six of those pages being distinct. At another site the user also visited 15 pages, with only five of those pages being distinct. In absolute terms, the scent at both Web sites appear to be poor since a number of previously visited pages were visited again. However, relative to one another, the first site appears to have a stronger scent then the second.

The following subsections illustrate manners in which the value of a patch and level of information scent can be gleaned from the clickstream of a user. By taking a user-centric viewpoint, many of the proposed conceptualizations are relative to the user's browsing behavior at other Web sites.

Table 8 lists the nine hypotheses of the user-centric model. The following subsections provide the rationale behind each of the hypotheses.

Table 8: User-centric: Hypotheses

Hypothesis #	Hypothesis
INFORMATION PATCH – SITE-PATCH	
UC1	Higher total duration spent at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.
UC2	Higher number of pages viewed at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.
UC3	Returning to this site-patch during the same user-session will be positively associated with achieving a goal on this long tail Web site.
UC4	Returning to this site-patch during a different user-session will be positively associated with achieving a goal on this long tail Web site.
INFORMATION PATCH – PAGE-PATCH	
UC5	Visitation of more highly valued goal page-patches at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site, where value is defined as the: (a) maximum value of any visited goal page-patch. (b) value from the last visited goal page-patch. (c) summation of values from all visited goal page-patches.
UC6	Higher median total duration spent within visited goal page-patches at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.
STRICT INFORMATION SCENT	
UC7	A lower proportion of repeatedly visited pages at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.
UC8	A more linear clickstream at this site-patch relative to other site-patches in this user-session will be positively associated with achieving a goal on this long tail Web site.
RELAXED INFORMATION SCENT	
UC9	Following of more highly valued goal scent trails at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site, where value is defined as the: (a) maximum value of any followed goal scent trail. (b) value from the last followed goal scent trail. (c) summation of values from all followed scent trails.

Information Patch

An information patch is a grouping of similar information, like a Web page or Web site (Pirolli, 2007). What a patch represents depends on the level of analysis being examined. At a high level, an entire Web site can be considered a patch. At a lower level, a Web page or set of Web pages may be considered a patch. The term *site-patch* is used to denote an entire Web site as a patch, while *page-patch* refers to an individual Web page or set of Web pages as a patch.

The first four hypotheses in this section examine how browsing behavior can lead to goal achievement by considering the Web site as a patch (i.e., site-patch). A benefit of taking a site-patch perspective is only coarse data on browsing behavior is required. The last two hypotheses in this section, however, take a more detailed viewpoint by focusing on specific pages or sets of pages being visited (i.e., page-patches). Although concentrating on page-patches requires finer-grained data, the lower level of analysis may tease out differences not seen at the site-patch level between goal- and non-goal-achieving foragers.

Site-patch

Since a forager has imperfect information and limited computational facilities, an optimal decision of how long to spend in a site-patch is unlikely. Instead, a forager is likely to employ satisficing (Pirolli, 2007; Simon, 1956), making a decision that satisfies a need (e.g., rate of information gain) at some specified level. When reading online texts for learning, satisficing is a commonly used technique (Reader and Payne, 2007). Using satisficing, a forager will continue to spend time reading pages on a Web site as long as information of value is being obtained. Therefore, a higher total duration spent at one site-patch relative to other site-patches can be associated with obtaining more information relevant to a user's information goal, which leads to Hypothesis UC1.

Hypothesis UC 1: *Higher total duration spent at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.*

Prior research has found mixed support for the association between absolute total duration and the achievement of a goal. A positive, negative, and insignificant association was found dependent on the task on one e-commerce Web site (Sismeiro and Bucklin, 2004). A positive and insignificant association was found using site-centric and user-centric data at another group of e-commerce

Web sites, respectively (Padmanabhan et al., 2001).

Each additional page visited represents a decision point where the user believed the value of continuing to browse at this site-patch was higher than what they expected to find elsewhere. In a similar vein as hypothesis UC1, a forager will continue to visit pages within a site-patch as long as information of interest is still being obtained. Therefore, more pages viewed at one site-patch relative to others can be associated with obtaining more information relevant to a user's information goal, which leads to Hypothesis UC2.

Hypothesis UC 2: *Higher number of pages viewed at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.*

Empirically, support has also been mixed for the association between absolute number of pages viewed and conversion. Prior research has found a positive association (Awad et al., 2006; Moe, 2003), no association (Chatterjee et al., 2003), and mixed association depending on the task (Sismeiro and Bucklin, 2004) or type of pages viewed (Van den Poel and Buckinx, 2005).

While foraging within a site-patch, a user forms a general opinion of the value of the Web site. When leaving one site-patch for another, a forager believes greater value may be found elsewhere. However, if a user returns shortly after leaving, the forager was unable to find a more valuable site-patch. Therefore, the site-patch of interest is more likely than other site-patches to contain the information necessary to fulfill the user's goal, which leads to Hypotheses UC3.

Hypothesis UC 3: *Returning to this site-patch during the same user-session will be positively associated with achieving a goal on this long tail Web site.*

When the span of time between visits is greater, returning to a site-patch demonstrates the positive evaluation of the site in two manners. First, the act of returning to a site indicates the forager originally valued the site-patch enough to remember its existence. Second, having a general recollection of the site and then returning also indicates the site-patch is expected to contain the information needed to fulfill the user's goal, which leads to Hypotheses UC4.

Hypothesis UC 4: *Returning to this site-patch during a different user-session will be positively associated with achieving a goal on this long tail Web site.*

Prior research has found positive, negative, and insignificant support depending on the task for

the association between returning to a Web site after a session has ended and achieving a goal (Sismeiro and Bucklin, 2004). As far as can be determined, the exit and return of a user during a user-session has not been examined in prior research.

Page-patch

As previously discussed, a page-patch consists of a Web page or set of Web pages that collectively provide information for an individual. However, certain page-patches may provide more useful information to a user than others. The identification of which page-patches are useful is likely to be similar amongst foragers with comparable goals. Patches predominately useful to goal-achieving foragers are known as *goal page-patches*, whereas *non-goal page-patches* are patches primarily of use to non-goal-achieving foragers.

A user who visits more highly valued goal page-patches is likely to have a goal similar to the goal-achieving foragers on that Web site. The value of a patch is considered in three different ways: maximum, most recent, and summation. Maximum value contends a highly valued patch visited at any point during a session is needed for the forager to judge the site favorably and thus consider achieving a goal. The value of the most recent (i.e., last visited patch) conjectures a goal is more likely to be achieved soon after visiting a highly valued patch. Finally, summation hypothesizes that the overall evaluation of the Web site, in terms of its valuable patches, affects the decision of a forager to achieve a goal or not.

In comparison to other Web sites visited during a user-session, a user who visits relatively more valuable goal page-patches at this Web site is more likely to achieve a goal, which leads to Hypothesis UC5.

Hypothesis UC 5: *Visitation of more highly valued goal page-patches at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site, where value is defined as the:*

(a) maximum value of any visited goal page-patch.

(b) value from the last visited goal page-patch.

(c) summation of values from all visited goal page-patches.

Positive, negative, and non-significant associations between specific pages and conversion have been found in prior research (Sismeiro and Bucklin, 2004). Differences between the types of pages visited and conversion rate have been also found at one e-commerce Web site (Moe, 2003). The actual relationship between types of pages viewed and conversion was found to be mixed at another e-commerce site (Van den Poel and Buckinx, 2005). As far as can be determined, relationships between groups of pages (of potentially different types) and conversion have not been examined in prior research.

The simple visitation of goal page-patches; however, does not provide a complete indication of how a forager actually processes a page or set of pages. For example, if a forager spends a very short amount of time in a goal page-patch it may signal the user did not fully recognize the value of the patch. A lack of recognition may be because of a poorly expressed information goal or simply a different information goal from previous goal-achieving foragers. Regardless, either reason would unlikely result in goal achievement at this Web site.

Similar to hypothesis UC1, a forager will continue to spend time reading pages within goal patches as long as information of value is being obtained. However, unlike hypothesis UC1 only the time spent on pages within already identified valuable goal page-patches is considered. Thus, a higher median total duration spent within goal page-patches at one site-patch relative to other site-patches can be associated with obtaining more information relevant to a user's information goal, which leads to hypothesis UC6[1].

Hypothesis UC 6: *Higher median total duration spent within visited goal page-patches at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.*

As far as can be determined, prior research has not specifically examined the association between amount of time spent on goal page-patches and goal achievement.

Information Scent

This section presents three hypotheses dealing with information scent. In the first two hypotheses, information scent is characterized by considering a user's entire session as a single monolithic

[1]Goal page-patches are unique to each site.

piece. In both these hypotheses a fairly strict definition of information scent is considered which views any inefficiencies in a user's clickstream (e.g., backtracking) as having poorer scent. The last hypothesis takes a more detailed viewpoint by looking at information scent among different fragments of a user's session. In this hypothesis a more relaxed characterization of information scent is used which recognizes that complex sessions may still be of high scent even in the presence of some inefficiencies.

Strict Information Scent

When a forager has a single well-defined goal in mind it would be expected the user would exhibit a focused search pattern (Moe, 2003). With a well-defined goal, the forager is better able to evaluate the scent of each link and hence make more accurate navigational choices. Viewed as a whole, such navigational choices for a forager with high levels of scent should result in a directed clickstream.

A directed path is characterized by few (if any) repeat visitations of pages, since it is assumed a rational forager would obtain any and all information from a page the first time it was visited. However, as even well-defined goals may be complex and hence result in less than direct clickstreams, scent relative to other Web sites visited is more appropriate to examine than absolute scent. Therefore, a goal is more likely to be achieved when a smaller proportion of pages are visited multiple times at this Web site relative to other sites, which leads to hypothesis UC7.

Hypothesis UC 7: *A lower proportion of repeatedly visited pages at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site.*

Empirically, the proportion of repeatedly visited pages has differed depending on the type of page and focus of the browser. For example, Moe (2003) found that directed shoppers at an e-commerce site viewed mostly unique product brand pages, somewhat unique category pages, and not very unique product pages. As far as can be determined, the use of proportion of repeated pages for an entire session has not been examined in prior research.

Taking a finer-grained conceptualization of strict information scent considers the overall complexity of a user's clickstream, as opposed to just general backtracking behavior. A less complex clickstream is one which exhibits a linear path through a site (Senecal et al., 2005), which is in-

dicative of high scent. As path information is used to determine complexity, backtracking behavior at many different pages rather than a single page may be teased out from a session.

For example, consider a user's browsing behavior at two Web sites. At one site seven pages were visited and four of those pages were unique. All of the non-unique pages were the home page which was used as the main hub for all the other pages being visited. At the other site the same number of total pages and unique pages were visited. At this Web site, however, each non-unique page was different from one another. Although the clickstreams from both Web sites have the same proportion of repeatedly visited pages, the clickstream from the second site is more linear and thus less complex than the second.

With high scent, a forager will exhibit a less complex and more linear clickstream than with low scent. However, similar to the previous hypothesis, absolute clickstream complexity is not appropriate to consider in light of potentially complex information goals. Therefore, a less complex clickstream, in terms of linearity, at this Web site relative to other Web sites is more likely to lead to goal achievement, which leads to Hypothesis UC8.

Hypothesis UC 8: *A more linear clickstream at this site-patch relative to other site-patches in this user-session will be positively associated with achieving a goal on this long tail Web site.*

Session complexity has been used to successfully discriminate users via their clickstream into high and low scoring groups (McEneaney, 2001); in the use of product recommendation agents (Senecal et al., 2005); and in predicting the completion of informational and e-commerce tasks (Kalczynski et al., 2006).

Relaxed Information Scent

The previous two hypotheses considered the session as a whole and assumed two things. First, inefficiencies in a user's clickstream were considered indicators of poor scent. However, certain "inefficiencies" may instead be a part of the natural decision making process of a user. For example, Moe (2003) found that when directed shoppers were deciding between products, their clickstreams demonstrated multiple repeated visits to the pages of the products being considered. Second, it was assumed the forager had a single information goal in mind when foraging. However, Montgomery et al. (2004) demonstrated that models which accounted for changes in visitors' goals on an e-commerce Web site performed better at predicting conversion than models which

only allowed for a single goal.

As the information goal of a forager may change during a session (or subgoals may be introduced as information is obtained) the scent of links will change in accordance to the current goal. Therefore, a link to a page which has already been visited might be selected again because (1) the link has the highest scent for the new current goal and (2) the scent gives an indication of novel information on the linked page. So even though the path at the aggregated session level of analysis may appear undirected due to a non-linear path or repeated viewings of the same page, if a forager's clickstream were separated by goal, a more directed manner of browsing within the context of the current information goal would likely be seen.

Figure 26 illustrates an example of an undirected path at the session level of analysis and a directed path at the goal level. The entire session consists of five page views with 60% of those pages being unique. Although the session as a whole does not appear to be directed, breaking the session down by the user's information goals reveals a different pattern. Within the context of a particular information goal, the pages viewed were unique as evidenced by the 100% path uniqueness for each goal's path subset.

Path Subset	Pages Viewed	Path Uniqueness
Entire Session	A, B, C, B, A	60%
Information Goal 1	A, B, C	100%
Information Goal 2	B, A	100%

Figure 26.: User-centric: Example Forager's Path

Thus, a finer-grained conceptualization of information scent is needed which is capable of detecting high scent in situations of changing information goals and "inefficient" behavior. To meet that need, *goal scent trails* and *non-goal scent trails*[2] are used in a similar spirit as goal and non-goal page-patches from hypothesis UC5. Goal scent trails are path fragments that goal-achieving foragers predominately follow. Non-goal scent trails are predominately followed by non-goal-achieving foragers.

[2]Olston and Chi (2003) introduced the concept of ScentTrails which highlighted the path a user should take given an information goal. ScentTrails differ from *scent trails* in that the former shows a path through a Web site given a user's goal, whereas the latter uses past foragers' behavior to determine goal and non-goal path fragments.

By only using portions of users' paths to derive scent trails, those parts of a session most and least aligned with goal-achieving can be teased from an entire session. A user who follows more highly valued goal scent trails is likely to have a goal similar to the goal-achieving foragers on that Web site. In the same manner as hypothesis UC5, the value of a scent trail is defined in three ways: maximum, most recent, and summation. Maximum value contends the most highly valued scent trail that is followed at any point during a session is needed for the forager to judge the site favorably and thus consider achieving a goal. The value of the most recent (i.e., last followed scent trail) conjectures a goal is more likely to be achieved soon after following a highly scented trail. Finally, summation hypothesizes that the overall evaluation of the Web site, in terms of its valuable trails, affects the decision of a forager to achieve a goal or not.

When compared to other Web sites visited during the same user-session, a forager who follows relatively more valuable goal scent trails at this Web site is more likely to achieve a goal, which leads to Hypothesis UC9.

Hypothesis UC 9: *Following of more highly valued goal scent trails at this site-patch relative to other site-patches within a user-session will be positively associated with achieving a goal on this long tail Web site, where value is defined as the:*

(a) maximum value of any followed goal scent trail.

(b) value from the last followed goal scent trail.

(c) summation of values from all followed scent trails.

Path information has been used successfully in clickstream research to predict future path selections (Montgomery et al., 2004). Various ways of representing paths have also been tested. The use of path fragments, which take into account the order, adjacency, and recency of information, have been found to be more accurate for predicting future paths than other manners of representing paths (Yang et al., 2004). As far as can be determined, the use of path fragments which distinguish between groups of a Web site population has not been examined in prior research.

Relation of Hypotheses to Information Foraging Theory

For each of the nine hypotheses, table 9 lists whether the hypothesis is testing or extending IFT. For each IFT-extending hypothesis, a short description is provided below which explains in what way the theory is being extended.

Table 9: Relation of Hypotheses to Information Foraging Theory

Hypothesis #	Hypothesis	Extends IFT?
INFORMATION PATCH – SITE-PATCH		
UC1	Duration	No
UC2	Number of pages	No
UC3	Leaving and returning	Partially
UC4	Returning back	Yes
INFORMATION PATCH – PAGE-PATCH		
UC5	Patch visitation	Yes
UC6	Patch duration	Partially
STRICT INFORMATION SCENT		
UC7	Unique pages	Yes
UC8	Linear clickstream	Yes
RELAXED INFORMATION SCENT		
UC9	Trail following	Yes

The first two hypotheses (UC1 – UC2) test IFT without extending the theory. Both of the hypotheses test the theory's expectation that users employ the concept of satisficing when foraging for information (Pirolli, 2007; Simon, 1956). As patches are assumed to exhibit diminishing returns, a visitor should only forage within a patch as long as they are satisfied with the rate of information gain they are obtaining.

The third hypothesis (hypothesis UC3) partially extends IFT. The idea is not novel that a forager would leave a patch when the rate of information gain falls below the mean rate of gain obtainable from the environment. However, the Marginal Value Theorem (Charnov, 1976) assumes an optimal forager with perfect information. Since foragers are known to possess imperfect information, the actual judgment on the mean rate of gain obtainable from other patches may be incorrect.

Therefore, a forager may return to the original patch after exploring other parts of the environment and realizing the original site still provided the highest rate of information gain.

Hypothesis UC4 is considered an extension to IFT because it introduces memory from past sessions. When searching for information, a forager will use information scent to guide them to patches of interest (e.g., a Web site). The level of scent recognized by a forager is dependent on the strength of chunks activated from declarative memory (Pirolli, 2007). It is assumed that when a forager visits a Web site of value, greater attention will be paid to the cues that represent that site compared to sites of a lower value. Greater cue attention will in turn more strongly activate the chunks representing those cues in declarative memory (Anderson et al., 2004). At a later time, when the forager has an information goal that may be achieved from the valuable Web site, those chunks representing the Web site will have a greater probability of being retrieved (than chunks representing lower-valued Web sites) from declarative memory due to being previously activated.

The hypotheses dealing with page-patches (UC5 – UC6) are also considered an extension because IFT does not define patches as being associated with a particular group of foragers (e.g., goal versus non-goal sessions). Instead, the patchy structure of the Web is assumed to be independent of a forager's information goal (Pirolli, 2007). Hypothesis UC5 is also an extension to the theory because patches in IFT are not given value independent of the current forager. Instead, the value of a patch is determined by an individual's behavior within that patch[3] (e.g., time spent).

The final three hypotheses are seen as an extension to IFT too. Within IFT, scent is viewed as a real-time mechanism that foragers use to select a navigational option (e.g., selecting which link to click next). While all three hypotheses still assume scent works by the same mechanism, an *overall* level of scent from a forager's *aggregated* behavior is conceptualized instead. In addition, hypothesis UC9 also extends IFT by introducing the concept of trails of scent that are common amongst foragers.

4.2.2 Site-centric

The site-centric model is useful when only the clickstream of a forager at a single site is known. As a result of having incomplete data; however, two ways in which concepts are defined to tap the main constructs of IFT in the user-centric model cannot to be used in the site-centric model.

[3]For example, value is equated with duration in hypotheses UC1 and UC6.

Instead, alternative forms of conceptualizing the constructs are needed.

The first way the definitions differ is in the usage of a forager's browsing behavior at the site of interest relative to their browsing behavior at the other sites visited during their user-session. Since the site-centric model has no knowledge of browsing behavior at other Web sites, comparisons are instead made relative to a fixed value of zero (i.e., in absolute terms)[4]. For example, users were assumed to have spent zero minutes and viewed zero pages on other Web sites. Thus, the site-centric model was reduced to only using a visitor's *absolute* browsing behavior at the site of interest.

The second difference between the two models is the ability to determine if the forager left the site and then came back during the session. Site-centric clickstream data would simply show a contiguous clickstream, regardless of if the forager left the site or not. However, site-centric datasets typically have access to a referring field which shows which URL a user came from (Fielding et al., 1999; Gourley and Totty, 2002). The use of the referring field is not without disadvantages as common browsing behaviors may lead the field to be blank (e.g., typing in a URL, using a bookmark). Despite these limitations, the use of referring information does provide a means that site-centric datasets may use to determine if foragers have left and returned to the site of interest within a session.

For example, figure 27 illustrates a site-centric view of the clickstream data available from a user. By looking at the user's entire clickstream (figure 25), it is known that the user left the site and returned after visiting page D1 the third time. But, the fact that the user left the site and returned cannot be determined from simply examining the site-centric clickstream as shown in figure 27. However, assuming the user followed links, the referring field would indicate page C1 was visited after the third D1 page and thus the forager left the site and returned.

With those two differences in mind, the hypotheses are restated for the site-centric clickstream model of information foraging in table 10.

[4]Chapter 8 provides a comparison of browsing behavior relative to users who had previously achieved a goal at the site of interest. The temporal version of the site-centric model assumes deviations from known goal-achieving browsing behavior indicates lower levels of scent or patch value and thus a lower probability of a goal being achieved.

Table 10: Site-centric: Hypotheses

Hypothesis #	Hypothesis
INFORMATION PATCH – SITE-PATCH	
SC1	Higher total duration spent at this site-patch will be positively associated with achieving a goal on this long tail Web site.
SC2	Higher number of pages viewed at this site-patch will be positively associated with achieving a goal on this long tail Web site.
SC3	Returning to this site-patch during the same session will be positively associated with achieving a goal on this long tail Web site.
SC4	Returning to this site-patch during a different session will be positively associated with achieving a goal on this long tail Web site.
INFORMATION PATCH – PAGE-PATCH	
SC5	Visitation of more highly valued goal page-patches will be positively associated with achieving a goal on this long tail Web site, where value is defined as the: (a) maximum value of any visited goal page-patch. (b) value from the last visited goal page-patch. (c) summation of values from all visited goal page-patches.
SC6	Higher median total duration spent within visited goal page-patches at this site-patch will be positively associated with achieving a goal on this long tail Web site.
STRICT INFORMATION SCENT	
SC7	A lower proportion of repeatedly visited pages at this site-patch will be positively associated with achieving a goal on this long tail Web site.
SC8	A more linear clickstream at this site-patch will be positively associated with achieving a goal on this long tail Web site.
RELAXED INFORMATION SCENT	
SC9	Following of more highly valued goal scent trails will be positively associated with achieving a goal on this long tail Web site, where value is defined as the: (a) maximum value of any followed goal scent trail. (b) value from the last followed goal scent trail. (c) summation of values from all followed scent trails.

The site-centric hypotheses have the same theoretical relation to IFT as the user-centric hypotheses. §4.2.1 provides an explanation of which hypotheses extend IFT and how the theory was extended.

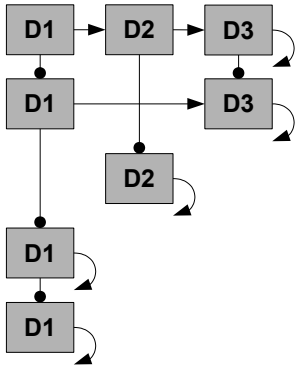


Figure 27.: Site-centric: Example User Clickstream Graph – Adapted from (Card et al., 2001)

4.3 Conclusion

This chapter provided an explanation on how IFT, a theory concerned with information search, can be used to help predict action. In addition, a brief overview was given of how an information forager processes a page and an example of the process using user-centric data. The user- and site-centric clickstream models of information foraging were then introduced. Finally, hypotheses generated from the user- and site-centric models were presented.

Chapter 5

Methodology

This chapter outlines the steps taken to test the hypotheses for both the user-centric (UC) and site-centric (SC) clickstream models of information foraging. The methodology for the user-centric model is presented first in §5.1, followed by the site-centric model in §5.2. For each model a description is given about the data sample and how to calculate each hypothesis' measure. Finally, §5.3 outlines the statistical tests used to test each hypothesis.

5.1 User-centric Clickstream Model of Information Foraging

The first subsection below describes how the user-centric dataset was processed to create user-sessions[1]. In the final subsection, details are given on how measures for the model's hypotheses were calculated. However, since only session-level information about a forager was available in the data, only measures which were calculable from the given data are presented (hypotheses UC1- UC4).

5.1.1 Dataset Sample

The user-centric dataset consists of a set of n sessions S (S_0, S_1, ..., S_{N-1}), where S_i represents a single session tuple. Each tuple consists of eight pieces of information: a unique identifier for the user, session, Web site, and referring domain; date and time the session started; number of pages viewed; how much time was spent on the site; and if the session resulted in a purchase being made (i.e., a goal). Table 11 illustrates a set of session tuples.

[1] Summary statistics about the user-centric dataset can be found in chapter 6.

Table 11: User-centric: Example Sessions

User	Session	Web site	Referring Domain	Date and Time	Pages	Duration	Goal?
U5	S1	W7	R3	5/25/08 15:39:07	12	17 min	Yes
U5	S2	W8	n/a	5/25/08 15:40:58	2	3 min	No
U6	S3	W5	n/a	5/25/08 15:53:02	5	9 min	No
U7	S4	W6	R1	5/25/08 16:02:34	3	4 min	No

User-sessions

The user-centric model is based on the idea of *user-sessions* which allow for an examination of a forager's behavior at one site relative to their behavior at other sites. A user-session U contains a target session T and a set of n other sessions S, where $S = \{S_0, S_1, \ldots, S_{n-1}\}$. T and S_i are both tuples that represents information about a particular session (as illustrated in table 11).

The target session T is a session that occurred at a long tail e-commerce site. In the dataset, a Web site was flagged as an e-commerce site if at least one purchase was made at the site by any user at any point during the dataset's time period. Web sites that made up the lowest 20% of all goals achieved were considered long tail e-commerce sites. A random sample of 20% of those long tail e-commerce Web sites with at least 50 goal sessions were selected for analysis[2]. Each session taking place at one of the selected long tail e-commerce sites became a target session for a potentially valid user-session.

To become a valid user-session, there must have been at least one other session at an e-commerce site by the user during the time the target session was active[3]. A session was considered active during the target session if it ended 30 minutes or less before the start of the target session[4]. In addition, the session must have also ended by the end of the target session[5]. At least one other session was required for a valid user-session in order to calculate relative behavior from the target

[2]Further details about the selection of long tail e-commerce sites may be found in §6.1.1.

[3]The other session could take place on *any* e-commerce site from the dataset.

[4]A 30 minute window before the beginning of the session was used because prior research has used a timeout period of 30 minutes for defining sessions (Bucklin and Sismeiro, 2003; Sismeiro and Bucklin, 2004; Van den Poel and Buckinx, 2005).

[5]The purpose of this research was to predict goal achievement at a particular instant in time (i.e., when the target session ended). Including sessions that ended after the target session would rely on data from the future. An entire session was removed from a user-session because the comScore dataset only included session-level information. Therefore, a session's browsing behavior could only be determined after a session had ended. If page-level information was available instead, the session's information known up to the target session's end would have been used.

session. Target sessions which did not have any other sessions during the window of time were not considered valid user-sessions and hence were not used in the analysis.

The *createUserSessions* algorithm in figure 28 illustrates the basic steps followed to create the user-sessions. The algorithm requires a set of long tail e-commerce Web sites and the number of minutes to use for a time window to be passed to the method when it is called. For each Web site, a set of sessions which visited the site were returned (line 19). These sessions were target sessions for potential user-sessions. For each target session, all other sessions by the user which (1) visited an e-commerce Web site and (2) ended between the specified number of minutes before the start of the target session or by the end of the target session were returned (line 22)[6]. If at least one other session was returned from line 22, then a user-session was created and added to the set of valid user-sessions (line 25). This process continued for each potential target session from each of the long tail e-commerce Web sites. After all processing was complete, a set of valid user-sessions was returned from the algorithm (line 29).

Table 12 illustrates how the *createUserSessions* algorithm operates. The table lists a subsample of sessions from the same user at e-commerce sites sorted by session date and time. The "Long Tail?" column specifies whether the site visited was a long tail e-commerce Web site. The final column, "Target?", specifies which long tail Web site was the focus of the user-session. The target column is provided to be clear which Web site would be used to compare against, especially in the situation of multiple long tail sites existing within the same user-session.

Assuming Web site $W5$ was currently being processed, then session $S4$ would have been returned from line 17 of the algorithm. Each of the returned sessions from the Web site would have then been iterated through. When session $S4$ was processed, sessions $S3$ and $S5$ would have been returned from line 20 of the algorithm. Session $S3$ would have been returned because the end of the session was within 30 minutes of the start of target session $S4$ (11:35:00 - 11:33:00 = 2:00). Session $S2$ would not have been included because the end of the session was more than 30 minutes from the start of the target session (11:35:00 - 11:04:00 = 31:00). Although session $S5$ started after the target session, it would still be included because the end of the session was equal to or less than the end of the target session (11:44:00 for both sessions).

Since two other sessions were found for the target session, a valid user-session would have been

[6]The target session was not returned in the set of other sessions in line 22 of the *createUserSessions* algorithm.

```
1 /**
2  * Parameters: (a) Set of long tail Web sites:
3  *                 W = {W_0, W_1, ..., W_{n-1}}
4  *             (b) Time window duration in minutes: timeWindow
5  * Returns:    Set of valid user-sessions:
6  *                 U = {U_0, U_1, ..., U_{m-1}}
7  *                 where U_i is a tuple:
8  *                 <T, {S_0, S_1, ..., S_{N-1}}>
9  * Methods:    (a) getSessions(w): returns set of sessions from Web site w
10 *             (b) getOtherSessions(s, time): returns set of
11 *                 session for this user from any e-commerce Web site
12 *                 within the specified window of time
13 *             (c) createUserSession(s, O): returns a valid user-session
14 */
15 createUserSessions(W, timeWindow) {
16     U = {};
17
18     for each (w ∈ W) {
19         S = getSessions(w);
20
21         for each (s ∈ S) {
22             O = getOtherSessions(s, timeWindow);
23
24             if (||O|| > 0) {
25                 U += createUserSession(s, O);
26             }
27         }
28     }
29     return U;
30 }
```

Figure 28.: User-centric: createUserSessions Algorithm

Table 12: User-centric: Example User-Sessions

User-session	Session	Web site	Date and Time	Duration	Long Tail?	Target?
	S1	W8	5/25/08 10:00:00	17 min	No	–
	S2	W4	5/25/08 11:00:00	4 min	No	–
U1	S3	W7	5/25/08 11:30:00	3 min	No	–
U1, U2	S4	W5	5/25/08 11:35:00	9 min	Yes	U1
U1, U2	S5	W6	5/25/08 11:40:00	4 min	Yes	U2
	S6	W2	5/25/08 13:00:00	19 min	No	–
U3	S7	W1	5/25/08 15:00:00	23 min	No	–
U3	S8	W3	5/25/08 15:30:00	31 min	Yes	U3
	S9	W2	5/25/08 18:05:00	23 min	Yes	U4

created on line 23 of the algorithm. The valid user-session would have included target session $S4$ and other sessions $S3$ and $S5$.

Table 12 also illustrates three other potential user-sessions. Both user-sessions $U2$ and $U3$ would have been valid because they both included other sessions beyond their target sessions (session $S4$ for $U2$ and $S7$ for $U3$). User-session $U4$ would not have been valid because there were not any other sessions within the time window of session $S9$. User-session $U4$ would not have been included in the analysis.

5.1.2 Metrics

Table 13 summarizes the metrics used to test the hypotheses for the user-centric clickstream model. The name of each metric along with a description of how it was calculated is provided. In addition, the hypothesis which corresponds to the metric is also provided in the table. A more in-depth description of the metrics is given in the following subsections.

Table 13: User-centric: Model Metrics

Hypothesis #	Metric	Description
INFORMATION PATCH – SITE-PATCH		
UC1	RELDUR	Duration in minutes spent on a Web site relative to median time spent during other sessions.
UC2	RELPGS	Number of pages viewed on a Web site relative to median number of pages from other sessions.
UC3	RETURN	If visitor left the Web site and returned during the same session.
UC4	VISITED	If visitor had previously visited the Web site before.
OTHER		
n/a	GOAL	Whether a goal occurred during the session.

To help clarify the notation being used below for the metrics, each user-session U contains a target session T and a set of n other sessions S, where $S = \{S_0, S_1, \ldots, S_{n-1}\}$. T and S_i are both tuples that represents information about a particular session (see §5.1.1 for more details).

Information Patch – Site-Patch

RELDUR is the total duration in minutes a visitor spent at the target Web site relative to the median time spent at other sites within the same user-session. The relative duration of the user-session U is calculated from equation 5.1, where $duration(i)$ is the duration spent during session i. To acquire RELDUR, the median duration of all sessions in the user-session U is subtracted from the total duration of the target session T.

$$\text{RELDUR} = duration(T) - median\left(\textbf{for each}_{i \in S}\left[duration(i)\right]\right) \tag{5.1}$$

RELPGS is the number of pages viewed at the target Web site relative to the median number of pages viewed at other sites within the same user-session. The relative number of pages for the target session T is calculated as shown in equation 5.2, where $pages(i)$ is the number of pages viewed during session i. To obtain RELPGS, the median number of pages viewed at the other Web sites is subtracted from the number of pages viewed during the target session.

$$\text{RELPGS} = pages(T) - median\left(\textbf{for each}_{i \in S}\left[pages(i)\right]\right) \tag{5.2}$$

RETURN is a binomial variable which is *true* if the user left and returned to the target Web site during the user-session and *false* otherwise. A user is designated as leaving and returning to the target Web site if another session is active during some part of the target session. This can occur if a new session is started while time is still being spent at the target Web site. Another situation where this can occur is if a session was started before the target session and continues to be active during some portion of the target session.

For example, RETURN would be *true* if session $S4$ from user-session $U1$ (table 12) was the target session. This is because session $S5$ started (11:40:00) during the time session $S4$ was still active (11:35:00 to 11:44:00). RETURN would be *false*, however, if session $S8$ from user-session $U3$ was the target session. Since session $S7$ was finished (15:00:00 to 15:23:00) before session $S8$ began (15:30:00), the forager could not have left and returned to the Web site from $S8$.

VISITED is a binomial variable which is *true* if the forager had visited the target Web site during another session at some point in the past and *false* otherwise. VISITED is calculated by examining the prior sessions of a forager and determining if the user had ever visited the Web site of interest

before.

Other

The mutually exclusive binomially distributed metric GOAL specifies whether a purchase was made during the session. If a goal was achieved during the session, GOAL would have the value of *true*. Otherwise, GOAL would have a value of *false*.

5.2 Site-centric Clickstream Model of Information Foraging

In the first subsection below, the methodology is presented on how the data was used to test the site-centric model[7]. The final subsection details how the measures for the site-centric hypotheses were calculated. Unlike the user-centric dataset, the site-centric dataset was at the page-level and thus each of the measures for the site-centric hypotheses was able to be calculated.

5.2.1 Dataset Sample

The supplied data contained a set of n sessions S (S_0, S_1, ..., S_{n-1}), where S_i represents a single session. Each session (S_i) contained a set of m page information tuples P (P_{i0}, P_{i1}, ..., P_{im-1}), where P_{ij} represents information about a particular page viewed during a session. Each page information tuple was made up of seven pieces of information: a unique identifier for the session, Web site, referring domain, and page viewed; date and time the page was viewed; how much time was spent on the page; and if the page represented a contact goal being achieved.

Table 14 illustrates a set of page tuples for session $S9$ at Web site $W4$. Of note is the right-censored nature of the site-centric data. The duration on the final page of the session is missing because it is not known when the next page was visited by this user (at this site or another).

Table 15 provides some basic statistics on the number of pages viewed and total duration of session $S9$. The first row of the table shows statistics using the entire session. However, only those parts of a session occurring *before* the achievement of a contact goal were used in the analysis. This truncation was done because the problem being investigated was the prediction of goal achievement during the *remainder* of a session. Thus, prediction was done from a point right before a form submission occurred.

[7] Summary statistics about the site-centric dataset can be found in chapter 6.

Table 14: Site-centric: Example Session Tuples

Session	Web site	Referrer	Page	Date and Time	Duration	Contact Goal
S9	W4	W6	A	5/25/08 15:37:02	32 s	n/a
S9	W4	W4	B	5/25/08 15:37:34	93 s	n/a
S9	W4	W9	C	5/25/08 15:39:07	111 s	n/a
S9	W4	W4	D	5/25/08 15:40:58	95 s	CG1
S9	W4	W4	A	5/25/08 15:42:33	9 s	n/a
S9	W4	W4	E	5/25/08 15:42:42	n/a	n/a

Table 15: Site-centric: Example Session Statistics by Contact Goal

	Pages	Duration
Entire session	6	340 s
Contact Goal 1	3	236 s
Contact Goal 2	6	340 s

To illustrate the truncation of a session, assume contact goal $CG1$ was being examined (represented as page D). For session $S9$, only activity on pages A, B, and C would be used (as illustrated in the second row of table 15). If contact goal $CG2$ were being examined instead (represented as page R), then the activity from the entire session would be used. This is because session $S9$ never visited the page representing the submission of a contact form for contact goal $CG2$. Thus, all pages of session $S9$ were usable since they occurred before the non-existent submission.

5.2.2 Metrics

Table 16 summarizes the metrics used to test the hypotheses for the site-centric clickstream model. The name of each metric along with a description of how it was calculated is provided. In addition, the hypothesis which corresponds to the metric is also provided in the table. A more in-depth description of the metrics is given in the following subsections.

To help clarify the notation being used below for the metrics, each session contains a set of m page information tuples P, where $P =< P_0, P_1, \ldots, P_{m-1} >$. P_j represents information about

Table 16: Site-centric: Model Metrics

Hypothesis #	Metric	Description
INFORMATION PATCH – SITE-PATCH		
SC1	SITEDUR	Duration in seconds spent on a Web site.
SC2	SITEPGS	Number of pages viewed on a Web site.
SC3	RETURN	If visitor left the Web site and returned during the same session.
SC4	VISITED	If user had previously visited the Web site before.
INFORMATION PATCH – PAGE-PATCH		
SC5a	PATCHMAX	Maximum value of any goal page-patch visited.
SC5b	PATCHLAST	Value of last goal page-patch visited.
SC5c	PATCHSUM	Total value of all goal page-patches visited.
SC6	PATCHDUR	Median duration in seconds spent in all goal page-patches.
STRICT INFORMATION SCENT		
SC7	UNIQUE	Percentage of unique pages viewed.
SC8	LINEAR	Linearity of clickstream.
RELAXED INFORMATION SCENT		
SC9a	TRAILMAX	Maximum value of any goal trail followed.
SC9b	TRAILLAST	Value of last goal trail followed.
SC9c	TRAILSUM	Total value of all goal trails followed.
OTHER		
n/a	GOAL	Whether a goal occurred during the session.

a particular page viewed during the session (see §5.2.1 for more details). P *only* contains pages which occurred *before* the contact form was submitted for the contact goal of interest.

Information Patch – Site-Patch

SITEDUR is the total duration in seconds a visitor has spent at a Web site. The total duration for the current visitor is calculated according to equation 5.3, where $time(i)$ is the time spent on the i^{th} page.

$$\text{SITEDUR} = \sum_{i \in P} time(i) \tag{5.3}$$

SITEPGS is the number of pages viewed during a session. The number of pages viewed during the current user's session is simply $\|P\|$ (equation 5.4).

$$\text{SITEPGS} = \|P\| \tag{5.4}$$

RETURN is a binomial metric which is *true* if the user left and returned to the Web site during the session and *false* otherwise. Since the dataset is site-centric, the determination of leaving and returning to a Web site cannot always be definitively determined. However, in many cases however the HTTP referer [*sic*] field (Fielding et al., 1999) contains information on what URL a forager was on before arriving at the current page. Thus, if the referring URL from any page viewed in a session (except for the first page viewed) is from a domain other than the current Web site, it can be concluded the user left the site and returned. The preceding rule does not apply to the first viewed page of a session since a forager cannot leave a Web site and return before a session has actually started.

To illustrate, in table 14 (§5.2.1) the referrer of the third page viewed (P3) was from a different domain than the current Web site (W9 versus W4). Therefore, the forager would have a RETURN value of *true* since the user left site W4, visited W9, and then returned to site W4. The fact that the first page viewed (P1) had a referring URL of a different Web site (W6 versus W4) has no bearing on the value of RETURN.

VISITED is a binomial metric which is *true* if the forager had visited the Web page during another session at some point in the past and *false* otherwise. VISITED is calculated by examining the

prior sessions of a forager and determining if the user ever visited the Web site of interest before.

Information Patch – Page-Patch

Patches at a Web site must already be known in order to calculate the four PATCH visitation metrics: PATCHMAX, PATCHLAST, PATCHSUM, and PATCHDUR. The methodology for learning patches is described in detail in appendix 5.B. In general, learning patches requires a set of goal and non-goal sessions to determine which parts of a Web site (i.e., pages) are better able to distinguish between the two groups. Patches are specific to a single Web site.

As the four PATCH metrics require patches to be learned first in order to quantify a session's patch visitation, the sessions for each Web site were split into two groups: training and testing sets. The training set was used to discover goal patches at a Web site. The sessions in the testing set each calculated the PATCH metrics for their individual session from the learned goal patches. However, a session from the testing set would only calculate the PATCH metrics *if and only if* goal patches were found at the Web site. In addition, the PATCHDUR metric would only be calculated for a session from the testing set *if and only if* that session visited at least one of the Web site's discovered goal patches.

Training and Testing Set

Each Web site contained sessions where either a goal was achieved during a session or not. Sessions were separated according to their achievement and placed into a Web site's goal dataset (D_G) or non-goal dataset (D_N)[8]. To create a Web site's training set (R), the sessions from both the goal and non-goal datasets were sorted in ascending order by their session start date. Then the first 70% of sessions from the goal dataset (D_G) were placed into the training set (R). The date of the last goal session added to R was noted. Sessions from the non-goal dataset (D_N) which occurred at or before the noted date of the last goal session from R were also added to the training set. All sessions from D_G and D_N not added to the training set were put into the testing set (E).

Learning Patches

Patches were learned for a Web site using the training dataset (R) according to the methodology outlined in appendix 5.B. Patches were learned at α levels of 0.05 and 0.01 and supported levels of

[8]To simplify notation, D_G and D_N are used to refer to the current Web site being examined.

0.25 to 1.50 (in 0.25 increments)[9].

Specifically, a set of n valuable patches A (A_0, A_1, ..., A_{n-1}) were discovered, where A_i represents a single valuable patch[10]. A_i consists of a set of m unordered and distinct pages U (U_0, U_1, ..., U_{m-1}).

Each patch (A_i) was also given a value according to equation 5.5 (Yang and Padmanabhan, 2003). S_{Gi} and S_{Ni} represent the number of goal and non-goal sessions from the training dataset that visited patch A_i, respectively. R_G and R_N is the total number of goal and non-goal sessions from the training dataset. The value of patch A_i could range from zero to two, with higher numbers representing a greater difference in support of the patch in distinguishing between goal and non-goal sessions (i.e., being more valuable).

$$value(A_i) = \frac{\left| \frac{S_{Gi}}{R_G} - \frac{S_{Ni}}{R_N} \right|}{\frac{1}{2}\left(\frac{S_{Gi}}{R_G} + \frac{S_{Ni}}{R_N} \right)} \tag{5.5}$$

Table 17 provides an example of three valuable patches found at a Web site. Each patch is made up of a set of unique and distinct pages. In addition, the value (as calculated from equation 5.5) is provided for each patch.

Table 17: Site-centric: Example Valuable Patches

Patch	Pages	Value
A1	{A, C}	0.75
A2	{B, C}	1.15
A3	{B, C, E}	1.35

Calculating PATCH Metrics

To calculate the PATCH metrics for a given session from the testing set (E), two steps were required. First, it was determined what patches the session visited from the set of valuable patches (A). Each session had a set of l visited patches V (V_0, V_1, ..., V_{l-1}), where V_j was an individual

[9] The results of the sensitivity analysis can be found in §7.2.3.

[10] A_i is a simplified form of notation which assumes a fixed Web site and significance or support level.

patch visited by the current session[11]. A session was considered to have visited a patch if all pages of the patch (U) were visited at least once (in any order) by the current session (as determined by the set of pages P from the session). Formally, A_i was added to V if $U \subseteq P$. Once it was known what patches were visited, then the four measures were calculated.

Table 18 provides an example of two patches visited by a session with three page views ($< A$, B, $C >$). $V1$ and $V2$ are simply patches $A1$ and $A2$ from table 17 that were visited by the session. $A3$ was not included because the session never visited page E. Table 18 is also used to calculate examples for each of the measures in this subsection.

Table 18: Site-centric:
Example Visited Patches

Patch	Pages	Value
V1	{A, C}	0.75
V2	{B, C}	1.15

PATCHMAX is the value of the most valuable patch visited by the current user. The maximum value is determined by iterating over every visited patch to find the one with the highest value (equation 5.6). If the user did not visit any patches then the value of PATCHMAX would be zero.

$$\text{PATCHMAX} = \begin{cases} max\,(\textbf{for each}_{j \in V}(value(V_j))) & \text{if } \|V\| > 0 \\ 0 & \text{else} \end{cases} \tag{5.6}$$

To illustrate equation 5.6, PATCHMAX would be 1.15 ($max(0.75, 1.15)$) assuming a user visited the patches in table 18.

PATCHLAST is the value of the last patch visited by the user. A four step heuristic was used to determine which patch was visited last during a user's session.

(1) For each patch visited, the position within the user's session when the forager *last* visited a page from that patch was noted[12]. PATCHLAST then equaled the value of the patch with the highest ending position. If more than one patch had the same highest ending position then the process continued to the second step.

[11] V_j is a simplified form of notation which assumes a valuable visited patch from a fixed Web site and significance or support level.

[12] If a user visited a page more than once, then the last time the page was visited was used.

(2) PATCHLAST equaled the value of the largest patch from the tied patches in the first step. Largest was defined as the patch with the highest number of pages (i.e., $\|U\|$). If more than one patch tied for the largest patch then the process continued to the third step.

(3) For each of the remaining tied patches from step two, the position within the user's session when the forager *first* visited a page from that patch was noted[13]. PATCHLAST then equaled the value of the patch with the highest starting position (i.e., started exploring the patch last). If more than one patch had the same highest starting position then the process continued to the fourth step.

(4) PATCHLAST equaled the median value of all tied patches from step three.

Table 19 illustrates the values obtained from following the heuristic on the visited patches from table 18. In this example, PATCHLAST would be 1.15. The steps for the heuristic for this example are provided after the table.

Table 19: Site-centric: Example Last Visited Patches

Patch	Pages	Value	Ending Position	Size	Starting Position
V1	{A, C}	0.75	3 ($max\,(1,3)$)	2	1 ($min\,(1,3)$)
V2	{B, C}	1.15	3 ($max\,(2,3)$)	2	2 ($min\,(2,3)$)

(1) The highest ending position for both patches was three. Since more than one patch was tied with the maximal value, a single patch could not be considered last, and thus the process continued to step two.

(2) Both patches also had a patch size of two. Therefore, the patches were tied again since neither of the patches was larger than the other patch.

(3) Patches $V1$ and $V2$ were first visited during the first and second page of the user's session, respectively. Since patch $V2$ had a later starting position it was deemed the last patch. Therefore, the value of PATCHLAST was the value of patch $V2$ (1.15).

[13] If a user visited a page more than once, then the first time the page was visited was used.

PATCHSUM adds up the value of every patch visited by the current user (equation 5.7). A value of zero is given to any user that did not visit any patches.

$$\text{PATCHSUM} = \begin{cases} \sum_{j \in V} (value(V_j)) & \text{if } \|V\| > 0 \\ 0 & \text{else} \end{cases} \tag{5.7}$$

PATCHSUM would be 1.90 ($0.75 + 1.15$) using the patches visited in table 18.

PATCHDUR is the median duration a user spent in all their visited patches. Only sessions which visited at least one patch (i.e., $\|V\| > 0$) would have a value for PATCHDUR. The calculation for PATCHDUR is shown in equation 5.8. $totalTime(k, P)$ returns the total time a session with pages P spent on page k. If a session visited page k more than once in P, then the sum duration from all k page visitations was returned.

$$\text{PATCHDUR} = median \left[\textbf{for each}_{j \in V} \left(\sum_{k \in G} totalTime(k, P) \right) \right] \tag{5.8}$$

PATCHDUR would be 164.50 s ($median(125, 204)$) for visited patches $V1$ and $V2$ (table 18) and session $S9$ (table 14).

Strict Information Scent

UNIQUE is the percentage of unique pages viewed during a session. The percentage of unique pages viewed for the current visitor is calculated according to equation 5.9, where $distinct(P)$ is the number of distinct pages viewed in the set of page information tuples P.

$$\text{UNIQUE} = \left(\frac{distinct(P)}{\|P\|} \right) * 100 \tag{5.9}$$

LINEAR is the complexity of a session as calculated via the stratum measure. Complexity is determined via the straightness (i.e., absence of visiting pages repeatedly) of a user's browsing behavior, where higher linearity equates to less complexity. Stratum is a measure of linearity from graph theory (McEneaney, 2001) and details on its calculation may be found in appendix 5.A.

Relaxed Information Scent

The three TRAIL metrics for the relaxed information scent were calculated in a very similar manner as the PATCH metrics. The same training set used to discover patches was used to learn trails. Sessions from the testing set then used those learned trails to calculate their values for the three TRAIL metrics.

Specifically, a set of n valuable trails T (T_0, T_1, ..., T_{n-1}) were discovered from the training set, where T_i represents a single valuable trail[14]. T_i consists of a set of m *ordered* pages O (O_0, O_1, ..., O_{m-1}), where the pages may repeat themselves in the ordered set (e.g., $< A, B, B, A, C >$). Once discovered, trails were given a value like patches using equation 5.5 (with T_i being used instead of A_i). Table 20 provides an example of three discovered trails.

Table 20: Site-centric: Example Valuable Trails

Trail	Pages	Value
T1	$< A, C >$	0.35
T2	$< A, A, C >$	1.25
T3	$< B, C, D >$	1.15

Once the trails were discovered, each session in the testing set (E) required two steps to calculate the TRAIL measures. First, it was determined what trails were followed by the session of interest from the set of valuable trails (T). Each session had a set of l followed trails F (F_0, F_1, ..., F_{l-1}), where F_j was an individual trail followed by the current session[15]. A session was considered to have followed a trail if all pages of the trail (O) were followed in order by the current session (as determined by the set of pages P from the session). Although all pages must have been followed in order, repeat visitation and gaps between pages were allowed (i.e., other pages may be visited in between pages from the trail). More specifically, T_i was added to F if $O \subseteq P$ and the pages of O were found in the same order in P. Once it was known what trails were followed, then the three measures were calculated.

Table 21 provides an example of two trails followed by a session with six page views ($<$ A,

[14] T_i is a simplified form of notation which assumes a fixed Web site and significance or support level.

[15] F_j is a simplified form of notation which assumes a valuable followed trail from a fixed Web site and significance or support level.

A, B, A, D, C $>$). $F1$ and $F2$ are simply trails $T1$ and $T2$ from table 20 that were followed by the session. $T3$ was not included because page C was not visited before page D in the session. Table 21 is also used to calculate examples for each of the measures in this subsection.

Table 21: Site-centric: Example Follwed Trails

Trail	Pages	Value
F1	$< A, C >$	0.35
F2	$< A, A, C >$	1.25

TRAILMAX is the value of the most valuable followed trail by the current user. The maximum value is determined by iterating over every followed trail to find the one with the highest value (equation 5.10). If the user did not visit any trails then the value of TRAILMAX would be zero.

$$\text{TRAILMAX} = \begin{cases} max\left(\textbf{for each}_{j \in F}(value(F_j))\right) & \text{if } \|F\| > 0 \\ 0 & \text{else} \end{cases} \tag{5.10}$$

To illustrate equation 5.10, TRAILMAX would be 1.25 ($max(0.35, 1.25)$) assuming a user followed the trails in table 21.

TRAILLAST is the value of the last trail followed by the user. A four step heuristic was used to determine which trail was followed last during a user's session.

(1) For each trail followed, the position within the user's session when the forager *last* visited the final page of the trail was noted[16]. TRAILLAST then equaled the value of the trail with the highest ending position. If more than one trail had the same highest ending position then the process continued to the second step.

(2) TRAILLAST equaled the value of the longest trail from the tied trails in the first step. Longest was defined as the trail with the highest number of pages (i.e., $\|O\|$). If more than one trail tied for the longest trail then the process continued to the third step.

(3) For each of the remaining tied trails from step two, the position within the user's session when

[16] If a user visited the final page of the trail more than once, then the last time the page was visited was used.

the forager *first* visited the first page of the trail was noted[17]. TRAILLAST then equaled the value of the trail with the highest starting position (i.e., started following the trail last). If more than one trail had the same highest starting position then the process continued to the fourth step.

(4) TRAILLAST equaled the median value of all tied trails from step three.

Table 22 illustrates the values obtained from following the heuristic on the followed trails from table 21. In this example, TRAILLAST would be 1.25. The steps for the heuristic for this example are provided after the table.

Table 22: Site-centric: Example Last Followed Trails

Trail	Pages	Value	Ending Position	Length	Starting Position
F1	$< A, C >$	0.35	6	2	1
F2	$< A, A, C >$	1.25	6	3	1

(1) The highest ending position for both trails was six. Since more than one trail was tied with the maximal value, a single trail could not be considered last, and thus the process continued to step two.

(2) Trail $F2$ had a length of three pages, while $F1$ only had two pages in its trail. Therefore, the value of TRAILLAST was the value of trail $F2$ (1.25).

TRAILSUM adds up the value of every followed trail by the current user (equation 5.11). A value of zero is given to any user that did not visit any trails.

$$\text{TRAILSUM} = \begin{cases} \sum_{j \in F} (value(F_j)) & \text{if } \|F\| > 0 \\ 0 & \text{else} \end{cases} \tag{5.11}$$

TRAILSUM would be 1.60 ($0.35 + 1.25$) using the trails visited in table 20.

[17] If a user visited the first page of the trail more than once, then the first time the page was visited was used.

Other

The mutually exclusive binomially distributed metric GOAL specifies whether at some point during the remainder of a session a contact form was submitted for the contact goal of interest. If a goal will be achieved during the session, GOAL will have the value of *true*. Otherwise, GOAL will have a value of *false*.

5.3 Metric Testing

Each of the metrics was tested individually to determine if they were able to distinguish between goal and non-goal sessions at *any* long tail Web site. The metrics were tested at the Web site unit of analysis since the goal was to find metrics which were significant over *multiple* long tail sites. Since each Web site had numerous goal and non-goal sessions, the median value was separately taken for each group[18]. The median values for the goal and non-goal sessions were then used as each Web site's paired data points.

The binomial metrics RETURN and VISITED did not use median values since the metrics were only flags indicating if someone left the site or had visited the site before. Therefore, each Web site was compared according to the probability of a goal occurring given if the user left and returned to the site or stayed at the site the entire session[19].

Table 23 illustrates the contingency table constructed for each Web site that was used to calculate the probabilities[20]. Counts of sessions at the Web site were categorized according to two dimensions: goal or non-goal session; and if the session left and returned or stayed on the site.

Equations 5.12 and 5.13 detail how each of the probabilities were calculated for each Web site. The probabilities were then used as each Web site's paired data points.

$$P\left(Goal|Return\right) = \frac{a}{a+b} \tag{5.12}$$

$$P\left(Goal|Stayed\right) = \frac{c}{c+d} \tag{5.13}$$

[18] Median values were used instead of mean values to reduce the impact of outliers on the dataset.

[19] For the VISITED metric the probabilities being compared were for a goal occurring given if the user had visited the site before or if this was the user's first visit to the site.

[20] All notations are stated for the RETURN metric. To be applicable to the VISITED metric, the notation "return" becomes "visited" and "stayed" changes to "new".

Table 23: Contingency Table for RETURN and VISITED

	Goal	Non-goal	N
Return	a	b	$a+b$
Stayed	c	d	$c+d$
N	$a+c$	$b+d$	$a+b+c+d$

A total of three different statistical tests were performed on each metric: paired t-test, exact Wilcoxon signed rank test, and dependent-samples sign-test. The paired t-test is a parametric test which assumes the data came from a normal distribution (Conover, 1999). The exact Wilcoxon signed rank test and the dependent-samples sign test are both non-parametric tests which do not make any assumption about the type of underlying distribution (Conover, 1999). The reason three tests were performed is due to each test's differing levels of assumption stringency. When metrics deviate from the assumptions of a test, the other less stringent tests can provide a "worst-case" baseline for the significance of the metric.

Each of the three tests is described in greater detail below, starting with the most stringent test. An example of each test is also provided which illustrates the test statistic being calculated.

Paired t-test

The paired t-test is a parametric test to determine if the mean difference between groups is zero (Conover, 1999). The difference of each pair's measures are calculated and then used to determine the test statistic t. The significance of t is then determined based on the assumption of an underlying normal distribution.

In the data there are n pairs of X and Y observations (X_0, Y_0), (X_1, Y_1), ..., (X_n, Y_n) (Conover, 1999). For each observation pair, the difference D_i is calculated between X_i and Y_i, where $D_i = Y_i - X_i$.

The test statistic t is calculated according to equation 5.14 (Conover, 1999), where $\bar{D}$ is the mean of all D_is.

$$t = \frac{\bar{D}}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} \left(D_i - \bar{D}\right)^2}} \tag{5.14}$$

Assumptions

The five assumptions for the paired t-test are provided below. The most stringent assumption for the test is the requirement of normally distributed random variables.

(1) "The D_is are identically distributed normal random variables." (Conover, 1999, pg. 363)

(2) "The distribution of each D_i is symmetric.

(3) The D_is are mutually independent.

(4) The D_is all have the same mean.

(5) The measurement scale of the D_is is at least interval." (Conover, 1999, pg. 353)

Example

To illustrate the paired t-test, table 24 provides an example of data from five Web sites (A-E). Within each Web site the median value for the metric being investigated is provided separately for the goal (X_i) and non-goal sessions (Y_i). In addition, the final column shows the difference (D_i) between the non-goal and goal sessions (i.e., $Y_i - X_i$).

Table 24: Example T-test Metric Testing Dataset

Web site	Goal Sessions (X_i)	Non-goal Sessions (Y_i)	D_i
A	3.75	2.15	1.60
B	7.15	4.35	2.80
C	12.20	13.40	−1.20
D	4.75	4.75	0.00
E	7.50	5.90	1.60

Using equation 5.14, the t-statistic for the given data is 1.3720 (with four degrees of freedom), and a p-value of 0.121 (assuming a hypothesis that X is greater than Y).

Exact Wilcoxon Signed Rank Test

The exact Wilcoxon Signed Rank Test (Wilcoxon, 1945) is a non-parametric test that determines if paired observations have the same mean as one another (Conover, 1999). Each Web site is ranked according to its absolute difference between the median values of all goal and non-goal sessions for the given metric. The ranks of all Web sites with a positive difference are then added up to obtain the test statistic V (Dalgaard, 2008).

In the data there are m pairs of X and Y observations (X_0, Y_0), (X_1, Y_1), ..., (X_m, Y_m) (Conover, 1999). For each observation pair, the difference D_i is calculated between X_i and Y_i, where $D_i = Y_i - X_i$. Any observation pair with a difference of zero is removed from the analysis (i.e., when $D_i = 0$). A total of n observation pairs then remain, where $n \leq m$.

The n remaining observation pairs are then ranked from 1 to n, where R_i is the rank of the i^{th} observation pair. Observations pairs are ranked from the smallest to the largest value of *absolute* difference (i.e., $|D_i|$). In cases where more than one observation pair shares the same *absolute* difference, then the assigned rank is averaged amongst all tied pairs. For example, assume the rank being assigned was three and four observation pairs shared the next smallest absolute difference. The rank for all four pairs would then be 4.5 ($\frac{3+4+5+6}{4}$).

After ranking, R_i takes the same sign as D_i (e.g., if D_i is negative then R_i is also negative). The ranks of all positive R_is are then summed to obtain the test statistic V (Dalgaard, 2008). An exact p-value is then computed from V using the Shift-Algorithm (Streitberg and Röhmel, 1986) in R (R Development Core Team, 2008).

Assumptions

The exact Wilcoxon signed rank test has the same assumptions as the t-test, except it does not require identically distributed normal random variables. The four assumptions for the Wilcoxon test are listed below.

(1) "The distribution of each D_i is symmetric.

(2) The D_is are mutually independent.

(3) The D_is all have the same mean.

(4) The measurement scale of the D_is is at least interval." (Conover, 1999, pg. 353)

Example

To illustrate the exact Wilcoxon signed rank test, table 25 provides an example of data from five Web sites (A-E). Within each Web site the median value for the metric being investigated is provided separately for the goal (X_i) and non-goal sessions (Y_i). The fourth column is the calculated difference (D_i) between the non-goal and goal sessions (i.e., $Y_i - X_i$). Since Web site D had a difference of 0.00, it was removed from any further analysis.

Table 25: Example Wilcoxon Metric Testing Dataset

Web site	Goal Sessions (X_i)	Non-goal Sessions (Y_i)	D_i	$\|R_i\|$	R_i
A	3.75	2.15	1.60	2.50	2.50
B	7.15	4.35	2.80	4.00	4.00
C	12.20	13.40	−1.20	1.00	−1.00
D	4.75	4.75	0.00	—	—
E	7.50	5.90	1.60	2.50	2.50

The fifth column shows the rankings of the four remaining Web sites according to the absolute value of D_i. Web site C was ranked first because it had the smallest value of $|D_i|$ (1.20). The next smallest value of $|D_i|$ was tied between Web site A and E (1.60). Both Web sites were given a rank of 2.50 ($\frac{2+3}{2}$). Finally, Web site B was given a rank of 4.00 since it had the largest value of $|D_i|$.

The final column displays the rankings of each Web site after taking into account the sign of D_i. Web site C's sign for R_i was switched to negative since D_i had a value less than zero. After adding all positively ranked Web sites, the test statistic (V) for this example was 9.00, with a p-value of 0.125 (assuming a hypothesis that X is greater than Y).

Dependent-samples Sign Test

The dependent-samples sign test is a non-parametric test that can also be used to test if there are differences between observations. Since the sign test has less stringent assumptions than many other non-parametric tests, it can be used it many more situations. For example, if the differences (D_is) between observations were not symmetrical in the exact Wilcoxon signed rank test, the sign test could be used as an alternative. However, the sign test is generally less powerful than other

non-parametric tests (Conover, 1999).

In the data there are m pairs of X and Y observations (X_0, Y_0), (X_1, Y_1), ..., (X_m, Y_m) (Conover, 1999). For each observation pair, a comparison is made. Assuming the purpose of the test is to determine if $X > Y$, then a pair is classified as "+" if $X_i > Y_i$, "-" if $X_i < Y_i$, or "0" if $X_i = Y_i$. All tied observation pairs (i.e., classified as "0") are discarded from further analysis, leaving a total of n observation pairs.

The test statistic S is calculated by counting the number of observation pairs classified as "+". The p-value is then computed from S using R (R Development Core Team, 2008).

Assumptions

The sign test has the least stringent assumptions of any of the tests discussed in this section. Thus, this test is useful in providing a way of testing metrics that can not meet the assumptions of the other tests. The sign test has the following three assumptions.

(1) "The bivariate random variables (X_i, Y_i) ... are mutually independent.

(2) The measurement scale is at least ordinal within each pair.

(3) The pairs (X_i, Y_i) are internally consistent, in that if P(+) > P(-) for one pair (X_i, Y_i), then P(+) > P(-) for all pairs." (Conover, 1999, pgs. 157-158)

Example

To illustrate the sign test, table 26 provides an example of data from five Web sites (A-E). Within each Web site the median value for the metric being investigated is provided separately for the goal (X_i) and non-goal sessions (Y_i). The final columns provides the classification for each Web site (e.g., "+", "-", "0"). Since Web site D was classified as "0", it was removed from any further analysis.

After counting all the Web sites classified as "+", the test statistic (S) for this example was 3.00, with a p-value of 0.3125 (assuming a hypothesis that X is greater than Y).

Table 26: Example Sign Test Metric Testing Dataset

Web site	Goal Sessions (X_i)	Non-goal Sessions (Y_i)	Classified
A	3.75	2.15	+
B	7.15	4.35	+
C	12.20	13.40	-
D	4.75	4.75	0
E	7.50	5.90	+

5.4 Conclusion

The preceding subsections presented the methodology for both the user- and site-centric clickstream models of information foraging. First, a description was provided for each model's data sample and how the measures for each hypothesis were calculated. Finally, the three statistical tests used to test each hypothesis were presented.

5.A Clickstream Complexity Appendix

The clickstream complexity metrics compactness and stratum were originally developed by Botafogo et al. (1992) to assist in the design of hypertext document collections (i.e., Web sites). The metrics were meant to quantify the complexity and connectedness of Web pages within a Web site. Compactness dealt with how well connected Web pages were to one another, where high compactness meant most pages had links to most other pages. Stratum was concerned with the degree of linearity in which Web pages must be read. High stratum occurred if a structured order existed in which Web pages must be read one after another.

McEneaney (2001) extended the work of Botafogo et al. (1992) by adapting the compactness and stratum metrics to be useful for quantifying users' paths. This section details how compactness and stratum can be calculated from a user's clickstream. Although only the stratum metric is used in this research, the compactness metric is explained for completeness. First, an example clickstream for two users is presented. Then the steps to convert a user's clickstream to a directed graph, path matrix, distance matrix, and finally converted distance matrix are explained. Finally, equations are presented to calculate compactness and stratum from the converted distance matrix.

5.A.1 Example Clickstreams

Figure 29 illustrates clickstreams for two separate visitors, V1 and V2, using a Web behavior graph. Both foragers visited seven pages with four of those pages being distinct. The path of the first visitor (V1) is $< P1, P2, P2, P3, P2, P4, P2 >$ whereas the path for the second visitor (V2) is $< P1, P2, P3, P4, P2, P3, P1 >$.

(a) Visitor V1

(b) Visitor V2

Figure 29.: Site-centric: Example Clickstream Web Graphs

Table 27 lists the compactness and stratum values for each of the visitors. Visitor V1 shows a moderately connected clickstream (compactness) because page P2 links to two distinct page and is linked from three distinct pages. The linearity of V1's clickstream (stratum) is moderate since the path taken does not end where it began. The second visitor (V2) has an even more densely connected clickstream than V1 since many of the pages are linked to more than one other page. In contrast to the first visitor, however, V2 has a much less linear clickstream. Although V2 appears to do very little backtracking to a single page, stratum is low since the path finished where it began[21].

Table 27: Site-centric: Example Visitor Clickstream Complexity Metrics

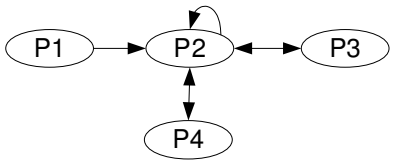
Visitor	Compactness	Stratum
V1	0.6389	0.6250
V2	0.7500	0.1250

[21]The value for stratum would change dramatically if visitor V2 would have visited page P4 instead of P1 as the last page of the path. With a path of $< P1, P2, P3, P4, P2, P3, P4 >$, the compactness and stratum values would be 0.5833 and 0.7500, respectively.

5.A.2 Graph Theory

Compactness and stratum are calculated by using concepts from graph theory (McEneaney, 2001). To graphically depict a clickstream, it can be converted into a directed graph. A directed graph consists of a set of nodes and directed links between the nodes. The nodes of a graph are the distinct Web pages viewed by a forager, while the links between nodes represent the transitions of a user from one page to another.

For example, figure 30a is a directed graph created from the first visitor's clickstream. The figure has four nodes representing each of the distinct pages visited. A single-headed arrow means the forager traveled from one node to another. A double-headed arrow represents a user traveling from one page to another and then back again. Of note is the sequence of the clickstream along with multiple traversals of the same path is lost when converting a clickstream to a directed graph.



(a) Directed Graph

To / From	P1	P2	P3	P4
P1	0	1	0	0
P2	0	1	1	1
P3	0	1	0	0
P4	0	1	0	0

(b) Path Matrix

To / From	P1	P2	P3	P4
P1	0	1	2	2
P2	∞	0	1	1
P3	∞	1	0	2
P4	∞	1	2	0

(c) Distance Matrix

To / From	P1	P2	P3	P4
P1	0	1	2	2
P2	4	0	1	1
P3	4	1	0	2
P4	4	1	2	0

(d) Converted Distance Matrix

Figure 30.: Site-centric: Example Clickstream Graph and Matrices

A way to represent the same information as the directed graph and allow for calculations is via a path matrix. A path matrix has each of the nodes as column and row headings. Each of the elements of a path matrix represents the number of transitions from one node to another. Initially, all elements in the matrix have values of zero. For each pair of nodes visited, the count at the intersection of those nodes in the matrix is increased. After processing all node pairs, any elements in the matrix with values greater than one are then set to one in order to create a "path adjacency matrix" (McEneaney, 2001, pg. 770).

Figure 30b illustrates the path matrix generated from the first visitor's clickstream. The value of one in the first row and second column represents the forager traveling from node P1 to node P2. In the next column over, the zero element means the user never went from node P3 to node P1. Figures 30a and 30b convey the same information in two different formats.

Using the path matrix, a distance matrix can then be created. The elements of a distance matrix represent the minimum distance (in terms of hops) between two nodes. The minimum distance between nodes is determined by using the shortest path algorithm by Floyd (1962). Unreachable paths between nodes are represented by the infinity symbol.

An example of a distance matrix from the first visitor's clickstream is shown in figure 30c. When going from node P1 to P3 there are two hops which must take place (P1 to P2 and P2 to P3), and thus the element has a value of two. Going from node P2 to P1 is set to infinity because only a path from node P1 to P2 exists, not one from node P2 to P1.

Since it is inconvenient to calculate the complexity metrics using infinite values, the distance matrix must be converted (Botafogo et al., 1992). Following Botafogo et al. (1992), all infinite values are replaced with the number of distinct nodes from the path matrix. In figure 30d the infinite element values are all replaced with the number four.

5.A.3 Compactness

Compactness is calculated according to equation 5.15 (McEneaney, 2001). $\|N\|$ is the number of distinct nodes in the user's clickstream. C is the converted distance matrix and C_{ij} refers to the element in the i^{th} row and j^{th} column. $\sum_i \sum_j C_{ij}$ simply sums all the elements from the converted distance matrix. The value of compactness ranges from zero to one, with values closer to one indicating a more densely connected and thus more complex clickstream (McEneaney, 2001).

$$\text{COMPACTNESS} = \frac{\|N\|^2 * (\|N\| - 1) - \sum_i \sum_j C_{ij}}{\|N\| * (\|N\| - 1)^2} \tag{5.15}$$

5.A.4 Stratum

Stratum is calculated according to equation 5.16 (Botafogo et al., 1992). AP and LAP both refer to equations more fully explained below. Values for stratum can range from zero to one, with values close to one indicating a more linear path and thus a less complex clickstream (McEneaney, 2001).

$$\text{STRATUM} = \frac{AP}{LAP} \tag{5.16}$$

Absolute prestige (AP) is the net status of a node within a hypertext network and is calculated according to equation 5.17 (Botafogo et al., 1992). S_i and CS_i refer to the status and contrastatus of a node. Status and contrastatus were originally developed for use in social network theory (SNT) (Harary, 1959). In SNT, status referred to the number of subordinates assigned to a person, whereas contrastatus was the number of superiors a person had. The same basic idea of status and contrastatus were adopted by Botafogo et al. (1992) for the stratum metric.

$$AP = \sum_i |S_i - CS_i| \tag{5.17}$$

The status of a node (S) shown in equation 5.18 is the number of other nodes which link from the node of interest (Senecal et al., 2005). Status is the sum of all non-infinite elements (e.g., $C_{ij} < \|N\|$) in a node's row from the converted distance matrix C.

$$S_i = \sum_i \begin{cases} \|N\| & \text{if } C_{ij} < \|N\| \\ \\ 0 & \text{otherwise} \end{cases} \tag{5.18}$$

Contrastatus (CS) is the number of nodes which link to the node of interest and is calculated according to equation 5.19 (Senecal et al., 2005). Contrastatus is the sum of all non-infinite elements (e.g., $C_{ij} < \|N\|$) in a node's column from the converted distance matrix C.

$$CS_j = \sum_j \begin{cases} \|N\| & \text{if } C_{ij} < \|N\| \\ \\ 0 & \text{otherwise} \end{cases} \tag{5.19}$$

Finally, equation 5.20 contains the formula for calculating the linear absolute prestige (LAN), which normalizes the size of the network for the stratum metric (Botafogo et al., 1992).

$$LAP = \begin{cases} \frac{\|N\|^3}{4} & \text{if } \|N\| \text{ is even} \\ \\ \frac{\|N\|^3 - \|N\|}{4} & \text{if } \|N\| \text{ is odd} \end{cases} \tag{5.20}$$

5.B Learning Patches and Scent Trails Appendix

This appendix is concerned with answering the first two research questions about how to learn information patches and scent trails[22]. Since the methodology for learning patches and scent trails is very similar to one another, the entire methodology is written in §5.B.1 from the viewpoint of learning patches. §5.B.2 provides a discussion of how the methodology differs for learning scent trails.

Research Question 1: *How can information patches be learned from a long tail Web site?*

Research Question 2: *How can information scent trails be learned from a long tail Web site?*

5.B.1 Information Patches

An information patch is defined as an area of the search environment with similar information (Pirolli, 2007). Within a Web-context, what constitutes a patch is dependent on the level of analysis being examined. At a high-level of analysis, an entire Web site can be considered a patch. When examined from a lower level of analysis, each individual page of a Web site can also be considered a patch. While such conceptualizations of a patch are straightforward, they are effectively being defined by the creator of the content rather than the user.

The Web, however, is a pliable environment where foragers have the choice of what material to view. Effectively, this allows a forager to define their own information patch that is uniquely relevant to their goal. Such patches may consist of a group of Web pages, which individually may mean very little, but when combined provide an area of the search environment that is seen as valuable to the user.

Although each user is free to define patches as they see fit, certain patterns of patches may emerge among foragers with similar information goals. From the viewpoint of the online firm, knowing who values what patch can provide insights into the information goal of the forager. By categorizing a patch as valuable to goal-achievers or non-goal-achievers, the firm may be able to better explain goal achievement at long tail sites dependent on what patches are visited by a user.

[22] Although mentioned together, the learning of information patches and scent trails are done separately from one another.

Learning Information Patches

An information patch is either a single Web page or a set of Web pages that collectively provide information for an individual[23]. From the perspective of the online firm, a patch is defined as valuable if it can distinguish between visitor sessions which result in a goal being achieved for the firm (e.g., a user purchases, or fills out a contact form), versus those that do not.

This section details how valuable information patches are learned. The first subsection provides the definition of a contrast set, which is used to discover patches. The methodology of learning patches using clickstream data and contrast sets is then outlined in the next subsection. The third subsection describes how patches are deemed to be valuable or not depending on their ability to significantly distinguish between goal-achievers and non-goal-achievers. Finally, an alternative definition of contrast sets is given, which does not require patches to be statistically significant between groups to be considered valuable.

Contrast Sets

From the data mining literature, contrast sets are a way to find differences between groups (Bay and Pazzani, 1999). A contrast set is a combination of attributes and their values which differ in support amongst separate groups (Bay and Pazzani, 1999). Let there be k attributes A (A_1, A_2, ..., A_k), where A_i can have one of m values (V_{i1}, V_{i2}, ..., V_{im}). A contrast set is a conjunction of attributes defined for n groups (G_1, G_2, ..., G_n) (Bay and Pazzani, 1999). For example, a contrast set may be $(PageA = 1) \wedge (PageC = 1)$, where the attributes represent Web pages and a value of "1" signifies a page was visited. Support in a group is defined as the percentage of instances where the contrast set is true within the group (Bay and Pazzani, 1999). The support from the previous example may be 5% for goal sessions and 17% for non-goal sessions.

A potential contrast set (PCS) is one where the contrast set ($cset$) is sufficiently large in at least one of the groups, where largeness is having a support greater than or equal to a specified minimum support ($minSup$). Formally, a PCS between two groups is one that satisfies the condition: $max(support(cset, G_1), support(cset, G_2)) \geq minSup$ (adapted from Satsangi and Zaiane (2007)). A significant contrast set (SCS) is a PCS that also meets the significance condition. Formally, a contrast set is significant between two groups if $P(cset|G_1) \neq P(cset|G_2)$ at a specified

[23]This appendix does not examine an entire Web site as a patch and thus the general term "patch" only refers to a single Web page or a group of Web pages.

alpha level (adapted from Bay and Pazzani (1999)). A SCS is hence a valuable information patch since it represents the fact that the set of pages tends to be visited more in one group than another. Therefore, the presence or absence of a visitor within such a patch may signal an expected goal outcome for the firm.

Discovering Patches

To discover valuable patches, contrast sets were found where the attributes of the set consisted of the distinct pages visited by a user during their session at a Web site. Certain pages, however, were not included in the analysis since their visitation was a requirement for or consequence of achieving a goal (e.g., contact form, form submission, and thank you pages). In addition, only pages occurring *before* a form submission were included in the analysis.

Each Web site contained sessions where either a goal was achieved during a session or not. Sessions were separated according to their achievement and placed into Web site i's goal dataset (D_{Gi}) or non-goal dataset (D_{Ni}). Each Web site had N_i sessions. N_{Gi} and N_{Ni} are denoted to correspond to the sizes of datasets D_{Gi} and D_{Ni} for Web site i, respectively.

Frequent itemsets were discovered from each Web site's datasets using the MAFIA (**Ma**ximal **F**requent **I**temsets **A**lgorithm) (Burdick et al., 2001) algorithm[24]. The algorithm was run separately on D_{Gi} and D_{Ni} for each Web site and resulted in a set of frequent itemsets I_{Gi} and I_{Ni}[25]. The minimum support was set to 0.10. A frequent itemset is a potential contrast set.

Figure 31 is an example of frequent itemsets mined from a Web site with three Web pages (A, B, and C) (assuming a $minSup$ of 10%). On the left-hand side of the figure are the itemsets discovered from the goal dataset (D_{Gi}), whereas the itemsets from the non-goal dataset (D_{Ni}) are on the right-hand side. The itemsets are arranged in a lattice by level according to their size (i.e., how many pages are in the itemset). Lines are drawn between itemsets to show their relation to other itemsets. To the right of each itemset in parentheses is the count of support for the itemset. The empty itemset at level 0 represents the entire dataset.

[24] An implementation of the algorithm can be found at http://himalaya-tools.sourceforge.net/Mafia/. Version 1.4 was used in this research.

[25] The discovery of frequent itemsets in datasets separated by goal is similar to discovering rules following the form $\{pages\} \rightarrow G$ as done in Satsangi and Zaiane (2007), where $\{pages\}$ is a set of distinct pages and G is the group goal or non-goal. However, when the size between groups is imbalanced, finding frequent itemsets in the minority group may become impossible using a combined dataset. For example, the minority group would not be able to find any frequent itemsets (at a minimum support of 10%) if the majority group had 10 times more records than the minority group. Mining frequent itemsets separately does not suffer from this class imbalance limitation.

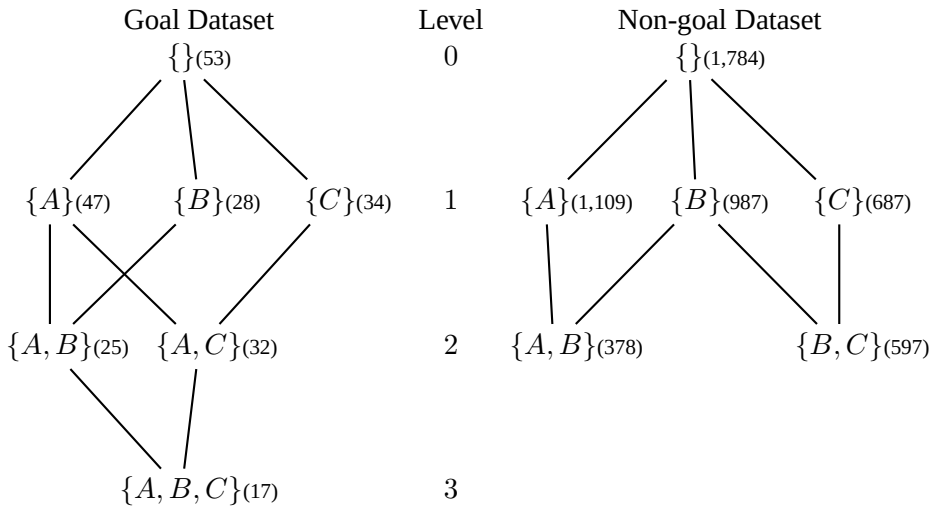


Figure 31.: Site-centric: Example Itemsets by Dataset

Potential contrast sets were formed starting with the lowest-level frequent itemsets found in either dataset and then continuing on to higher-level itemsets. To evaluate a PCS, a contingency table was needed that was populated with the amount of support and non-support for the PCS's itemset from each dataset. When the itemset for a PCS was found in both I_{Gi} and I_{Ni}, then the contingency table was created according to table 28, where S_{Gij} (S_{Nij}) is the count of support for itemset j from Web site i in the goal (non-goal) dataset.

Table 28: Site-centric: Example Contingency Table for a Potential Contrast Set

	Support Count for Itemset j	Non Support Count for Itemset j
D_{Gi}	S_{Gij}	$\neg S_{Gij} = N_{Gi} - S_{Gij}$
D_{Ni}	S_{Nij}	$\neg S_{Nij} = N_{Ni} - S_{Nij}$

When the itemset was missing from one of the datasets (i.e., it was not frequent), then the count of support and non-support was unknown. In such a case the support frequency for the contingency table (S_{Gij} or S_{Nij}) was calculated (Satsangi and Zaiane, 2007) according to the supCount formula: $supCount = round(N * minSup)$, where N is N_{Gi} or N_{Ni} and $minSup$ is minimum

support.

$supCount$ represents a generous count of support for an itemset, which results in a PCS being conservatively evaluated. When an itemset is not frequent it is unknown how much less than $minSup$ its support really is. Using a support count of zero in place of $supCount$ would underestimate the importance of an itemset, which may lead to a Type I error when evaluating the PCS. Therefore, a support count equivalent to $minSup$ is used which over-estimates the importance of an itemset and hence lowers the chance of a Type I error occurring when the PCS is evaluated. However, this method does increase the probability of a Type II error occurring.

To illustrate, assume $minSup$ was 10% and the support for a PCS's itemset was 9.9% in the goal dataset and 45.0% in the non-goal dataset. Since minimum support for the itemset from the goal dataset was not met, the true support would be unknown. If a support of zero were used when populating the contingency table for the PCS, the importance of the itemset in the goal dataset would be understated by 9.9%. In other words the difference in support between datasets would be 9.9% *more* than it was actually was (45.0% versus 35.1%). Setting the support to $minSup$ would instead over-estimate the importance of the itemset by 0.1%. Here the difference in support would be 0.1% less than it actually was (35.0% versus 35.1%).

Table 29 presents three examples of potential contrast sets from the second level of figure 31. The first column shows the itemset used for the PCS (i.e., the Web pages that make up the patch). Columns two through five are in reference to the goal dataset and list if the itemset was found to be frequent, number of goal sessions, support for the itemset (with support percentage in parentheses), and non-support for the itemset. Columns six through nine have the same meaning as columns two through five except they refer to the non-goal dataset.

Table 29: Site-centric: Example Potential Contrast Sets

	Goal				Non-Goal			
PCS	Found?	N_{Gi}	S_{Gij}	$\neg S_{Gij}$	Found?	N_{Ni}	S_{Nij}	$\neg S_{Nij}$
{A,B}	Yes	53	25 (47.2%)	28	Yes	1,784	378 (21.2%)	1,406
{A,C}	Yes	53	32 (60.4%)	21	No	1,784	178 (10.0%)	1,606
{B,C}	No	53	5 (9.4%)	48	Yes	1,784	597 (33.5%)	1,187

The first example shows a PCS for itemset $\{A, B\}$. Since the itemset was found in both datasets

the support from each dataset is known. The second PCS (itemset $\{A, C\}$) illustrates an example where the itemset of interest is found in the goal dataset but not in the non-goal dataset. Therefore, the $supCount$ formula was used (with a $minSup$ of 10%) to calculate the support count (S_{Nij}). The final PCS (itemset $\{B, C\}$) shows the opposite situation where the itemset was not frequent in the goal dataset but it was in the non-goal dataset. S_{Gij} would therefore be calculated using the $supCount$ formula[26].

Determining Patch Value

The significance of each potential contrast set was then calculated using Fisher's exact test (Conover, 1999). Although prior research has used the chi-square test for independence to determine significance (Bay and Pazzani, 1999), the approximation of α may suffer when the expected value of at least 20% of the cells in the contingency table are below five or any one expected value is less than one (Cochran, 1954). When considering goal achievement on long tail Web sites, the counts in the contingency table are often too small or too imbalanced in their distribution for the chi-square test to adequately approximate α. Thus, Fisher's exact test, which makes no such approximation, was used instead.

When testing multiple hypotheses, such as in the situation of testing each potential contrast set, the familywise error rate (FWER) should be controlled. The FWER is a measure in statistics that refers to the probability of committing at least one Type I error. A common method of dealing with the FWER is to fix the alpha across all tests.

For example, with an expected familywise error rate of α, the alpha level for each individual potential contrast set (α_{ind}) would be fixed using a Bonferroni procedure: $\alpha_{ind} = \alpha/NC$, where NC is the number of PCSs being tested. The disadvantage of such an approach is the same alpha-level is used regardless of the PCS's itemset size. This results in a loss of power and ability to detect differences in even the most general PCSs which use lower-level itemsets (Bay and Pazzani, 1999).

To combat such a loss of power, a different alpha level was used for each level of the itemset lattice. The purpose of such a change in α was to distribute "... 1/2 of the total α to tests at level 1, 1/4 to tests at level 2, and so on" (Bay and Pazzani, 1999, pg. 304). This results in greater power being available to test the most general PCSs (i.e., those with itemsets from the lowest levels).

[26]The support percentage is 9.4% instead of 10.0% due to rounding in the $supCount$ formula.

Equation 5.21 (Bay and Pazzani, 1999) was used to determine the alpha level (α_l) for testing all PCSs at a specified level. In the equation, α is the expected familywise error rate, l is the level, and C_l is the number of candidate PCSs being tested at level l. The purpose of the min function was to ensure that α becomes more stringent with each subsequent level. Using an α of 0.01, a potential contrast set was deemed significant if its p-value from Fisher's exact test was less than or equal to α_l.

$$\alpha_l = min\left(\frac{\frac{\alpha}{2^l}}{C_l}, \alpha_{l-1}\right) \tag{5.21}$$

If a PCS was found to be significant, then the patch it represents (i.e., itemset) was deemed valuable. A valuable patch which was predominately visited by users from the goal group was known as a *goal patch* and placed in the set P_{Gi}, whereas visitation mostly from the non-goal group resulted in a patch being labeled as a *non-goal patch* and being placed in the set P_{Ni}. Formally, a patch was deemed a *goal patch* if $\frac{S_{Gij}}{N_{Gi}} > \frac{S_{Nij}}{N_{Ni}}$, and a *non-goal patch* otherwise[27]. By classifying patches in such a manner, a visitor may signal expected goal outcome to the firm via the presence or absence of patch visitation.

Supported Contrast Sets

As an alternative to significant contrast sets, a supported contrast set was a potential contrast set that had a difference in support above a user-defined threshold ($thresh$). Formally, a supported contrast set between two groups was one that satisfies the condition:

$difference(support(cset, G_1), support(cset, G_2)) \geq thresh$, where $difference(S_{Gij}, S_{Nij})$ is defined in equation 5.22 (Yang and Padmanabhan, 2003). If a PCS met or exceeded the threshold support condition then the patch was considered valuable. The classification as either a goal or non-goal patch was done in the same manner as with significant contrast sets.

$$difference(S_{Gij}, S_{Nij}) = \frac{\left|\frac{S_{Gij}}{N_{Gi}} - \frac{S_{Nij}}{N_{Ni}}\right|}{\frac{1}{2}\left(\frac{S_{Gij}}{N_{Gi}} + \frac{S_{Nij}}{N_{Ni}}\right)} \tag{5.22}$$

The purpose of defining a supported contrast set was because finding statistical significance may be difficult when many PCSs exist. For example, assume 100 potential contrast sets existed

[27] Only goal patches were used in the analysis of this dissertation.

on level 1 and the expected familywise error rate was set to 0.01. In order for a potential contrast set to be significant, its p-value would need to be lower than 0.00005. Therefore, the supported definition of a contrast set was used to discover contrast sets which may be important, but fail to reach significance.

5.B.2 Scent Trails

Information scent is the driving force behind why a person makes a navigational selection amongst a group of competing options. As foragers are assumed to be rational, scent is a mechanism by which foragers' reduce their search costs by increasing their accuracy on which option leads to the information of value (Pirolli, 2007). Based on the information goal of a forager, each hyperlink on a Web page gives off a scent. The higher the scent the more likely the page that is being linked to may contain the information being sought. Similar to a bloodhound that follows a scent trail over distances to find an item of interest, a forager also follows a scent trail to find the information they seek over multiple Web pages.

Although each user follows a scent trail that fits with their information goal, patterns from fragments of scent trails may exist that emerge among foragers with similar information goals. Like patches, these fragments of scent trails are of value to the online firm in distinguishing between possible goal-achievers and non-goal-achievers. When a user follows these known fragments of scent trails it may provide clues into their information goal and thus help in explaining goal achievement at long tail sites.

Learning Scent Trails

A scent trail is the path a forager travels upon by following the information scent of links. More specifically, a scent trail is a set of pages in a specified order. To discover valuable scent trails, contrast sets were found where the attributes of the set consisted of an ordered set of pages (i.e., a sequential pattern) visited by a user during their session at a Web site.

Frequent sequential patterns were discovered from each Web site's datasets using the SPAM (**S**equential **PA**ttern **M**ining) algorithm (Ayres et al., 2002)[28]. The algorithm was run separately on

[28] An implementation of the algorithm can be found at http://himalaya-tools.sourceforge.net/Spam/. Version 1.3.3 was used in this research.

D_{Gi} and D_{Ni} for each Web site and resulted in a set of frequent sequential patterns P_{Gi} and P_{Ni}. A frequent sequential pattern is a potential contrast set.

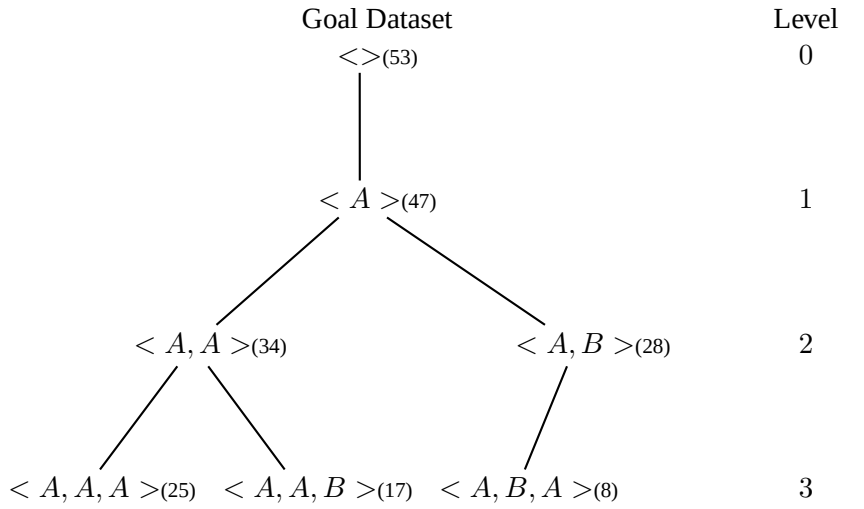
Figure 32 is an example of frequent sequential patterns mined from a Web site with two Web pages (A and B) (assuming a $minSup$ of 10%). The figure only shows patterns found frequent starting with Web page A. On the top part of the figure are the patterns discovered from the goal dataset (D_{Gi}), whereas the patterns from the non-goal dataset (D_{Ni}) are on the bottom part of the figure. The patterns are arranged in a lattice by level according to their size (i.e., how many Web pages are in the pattern). Lines are drawn between patterns to show their relation to other patterns. To the right of each pattern in parentheses is the count of support for the pattern. The empty pattern at level 0 represents the entire dataset.

An example of a potential contrast set with a sequential pattern from the third level of figure 32 is $< A,A,B >$. This pattern means page A was visited two times before page B was visited. However, the visitation of these pages need not be right after one another. There may be many other pages that were visited in between each page. For example, a session with the following seven page views $< C,A,C,D,E,A,B >$ visited pages C, D, and E before visiting page A again and then page B.

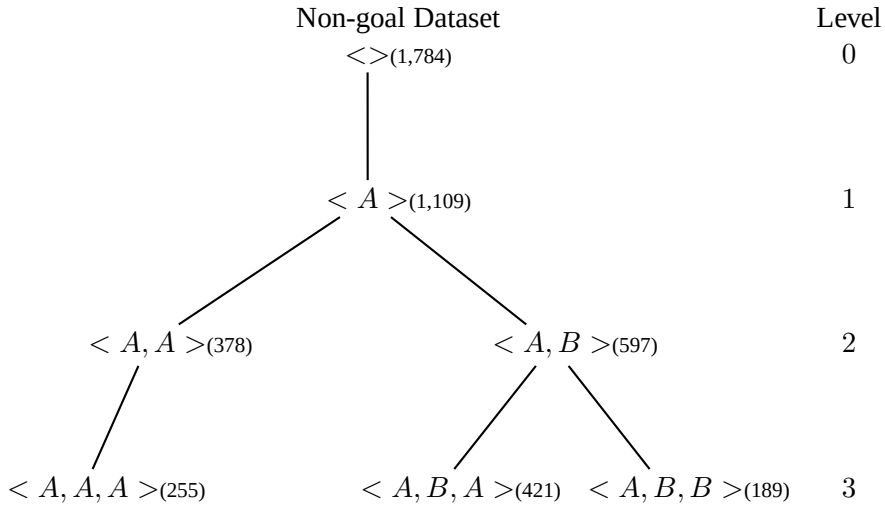
If a PCS was found to be significant, then the scent trail it represented (i.e., sequential pattern) was deemed valuable. A valuable scent trail which was predominately followed by users from the goal group was known as a *goal scent trail* and placed in the set T_{Gi}, whereas following mostly from the non-goal group resulted in a scent trail being labeled as a *non-goal scent trail* and being placed in the set T_{Ni}. Formally, a scent trail was deemed a *goal scent trail* if $\frac{S_{Gij}}{N_{Gi}} > \frac{S_{Nij}}{N_{Ni}}$, and a *non-goal scent trail* otherwise[29].

For a scent trail to be considered valuable by way of support (i.e., a supported contrast set), then the PCS must have met or exceeded the threshold support condition. The classification as either a goal or non-goal scent trail was done in the same manner as with significant contrast sets.

[29] Only goal scent trails were used in the analysis of this dissertation.



(a) Frequent Sequential Patterns from Goal Dataset



(b) Frequent Sequential Patterns from Non-Goal Dataset

Figure 32.: Site-centric: Example Patterns by Dataset – adapted from Ayres et al. (2002)

Chapter 6

Datasets

The datasets used to test the user- and site-centric clickstream models of information foraging are presented in this chapter. In particular, details of the preprocessing steps undertaken to arrive at the final dataset used to test each model are shown. After all the preprocessing steps, descriptive statistics of each final dataset are also provided. §6.1 contains information about the user-centric dataset, while §6.2 details the data from the site-centric dataset.

6.1 User-centric Dataset

The data used for the user-centric clickstream model of information foraging was provided by comScore, Inc., a marketing research company. The domain-level data was captured for 100,000 United States-based panelists[1] over a one year period from January 1, 2006 to December 31, 2006[2] (comScore, Inc., 2007b). Each panelist was randomly selected from comScore's pool of more than two million global Internet users.

Data was collected from panelists using a proprietary methodology that "...enable[d] comScore to passively observe the full details of panelists' Internet activity, including every Web site visited and item purchased" (comScore, Inc., 2005, pg. 1). A panelist's session was defined as any sequence of Web pages on the same Web site by the same visitor with less than a 30 minute time period between page viewings. A 30 minute session timeout has also been used in previous clickstream research (Bucklin and Sismeiro, 2003; Sismeiro and Bucklin, 2004; Van den Poel and Buckinx, 2005).

The remainder of this section provides information regarding the steps taken to arrive at the final dataset used to test the user-centric model, along with general descriptive statistics about the

[1] The actual dataset contained a total of 88,814 panelists. The documentation by comScore did not provide an explanation for the 11,186 missing panelists.

[2] Not all panelists were active during the entire data collection period.

data. The preprocessing steps applied to the data are described in the next section. Descriptive statistics are then provided in the following section.

6.1.1 Preprocessing of Original Dataset

The data obtained from comScore, Inc. included many data elements not applicable to the current research. Therefore, a number of processing steps were performed to obtain a final dataset usable for testing the user-centric clickstream model of information foraging. Table 30 lists each step of the process along with the total number of Web sites, sessions, and goal sessions; and how many Web sites, sessions, and goal sessions were removed at that step (if applicable). Table 31 lists the parameters used in each preprocessing step. A discussion of each step and its parameters are provided below.

Table 30: User-centric: Preprocessing of Original Dataset Statistics

Step	Description	Web Sites	Sessions[a]	Goals[b]	△ in Web Sites	△ in Sessions[a]	△ in Goals
	Original dataset	1,417,745	548,428,562	354,985	n/a	n/a	n/a
1	Remove non e-commerce sites	625	106,686,274	354,985	−1,417,120	−441,742,288	n/a
2	Remove sites not randomly selected from long tail	58	798,306	13,872	−567	−105,887,968	−341,113
3	Remove single page sessions	58	616,607	13,870	0	−181,699	−2
4	Identify user-sessions	58	511,397	11,100	n/a	−105,210	−2,770
5	Remove sites with < 50 goal user-sessions	52	502,131	10,834	−6	−9,266	−266
6	Remove outliers	52	496,343	10,714	n/a	−5,788	−120
7	Remove sites with < 50 goal user-sessions	52	496,343	10,714	0	0	0

[a] Starting at step 4, the "Sessions" and "△ in Sessions" columns refer to user-sessions.

[b] The original dataset contained a total of 355,064 goals. However, 79 of those goals (0.02%) did not have any session details associated with the purchase. Therefore, the total number of goals for this dataset was listed as 354,985.

Table 31: User-centric: Preprocessing Parameters

Step	Description	Parameters
	Original dataset	n/a
1	Remove non e-commerce sites	$goalsAtWebsite == 0$
2	Remove sites not randomly selected from long tail	$shortHeadWebsites == 80/20\ rule$ $minGoalsInLongTail == 50$ $randomPercent = 20\%$
3	Remove single page sessions	$sessionLength == 1$
4	Identify user-sessions	$sessionsInWindow \geq 1$
5	Remove sites with < 50 goal user-sessions	$userSessionGoalsAtWebsite < 50$
6	Remove outliers	$MinPts = 4$ $Eps = 0.0266$ (goal sessions) $Eps = 0.0536$ (non-goal sessions) $sample\% = 100\%$ (goal sessions) $sample\% = 15\%$ (non-goal sessions)
7	Remove sites with < 50 goal user-sessions	$userSessionGoalsAtWebsite < 50$

Step 1. Remove Non e-Commerce Web sites

The first step of the process removed any non e-commerce sites. An e-commerce Web site was defined as any site in which a purchase was made by any user at any point within the dataset's time period. A total of 625 Web sites were found where a purchase was made[3]. The remaining 1,417,120 non e-commerce Web sites were removed along with the 441,742,288 corresponding sessions that took place on those sites.

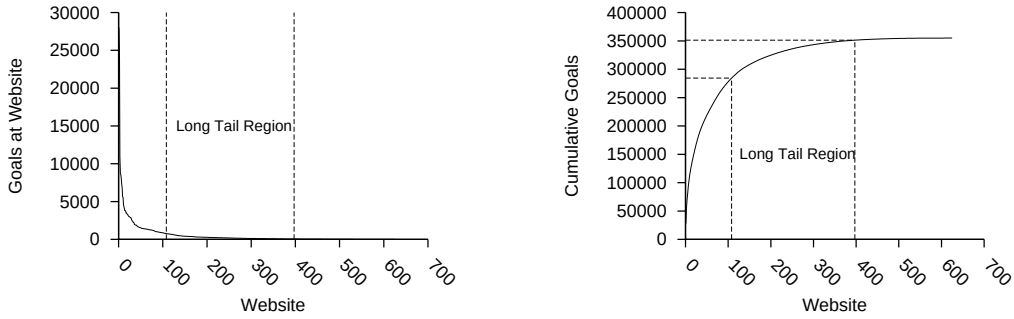
Step 2. Remove Web sites Not Randomly Selected from the Long Tail

The second step of the process selected a sample of long tail e-commerce Web sites to analyze. Sites within the dataset were defined according to the 80/20 rule (Newman, 2005) as either parts

[3]The methodology by which comScore recognizes and records a purchase was not available. Considering only a total of 625 out of 1,417,745 Web sites had purchases on them, it is likely the dataset did not include all purchases made at all Web sites.

of the short head or long tail of a power law distribution. In general, the 80/20 rule states 80% of some quantifiable object (e.g., wealth) should be held by 20% of the population. Within the context of goal achievement, 80% of achieved goals should have taken place on 20% (125) of the 625 e-commerce Web sites. Short head Web sites were those sites included in the 80/20 group, while all other sites were considered long tail Web sites.

Figures 33a – 33b illustrate the separation between short head and long tail Web sites. Figure 33a shows the number of goals achieved at each Web site, while figure 33b shows the cumulative number of goals achieved. The vertical dashed line on the left of each figure represents the boundary between short head and long tail Web sites. The Web sites to the left of the first dashed line represent 79.89% of all goal sessions, while making up 17.12% of the Web site population.



(a) Goal Sessions by Web Sites

(b) Cumulative Goal Sessions by Web Sites

Figure 33.: User-centric: Goal Sessions by Web Sites

The second vertical dashed line in each figure represents a separation between Web sites in the long tail and those in the very long tail. A Web sites was considered too far down the long tail to analyze if there were fewer than 50 goal sessions at the Web site.

A total of 107 Web sites (17.12%) in the short head were removed along with the 41,708,093 corresponding sessions (283,609 goals) at those Web sites. In addition, 228 Web sites (36.48%) in the very long tail were removed along with the 38,947,086 corresponding sessions (3,693 goals) at those Web sites.

290 Web sites (46.40%) in the long tail region remained with 26,031,095 sessions (67,683 goals). Since processing over 26 million sessions would be too computationally expensive, a random sample of 20% of the 290 Web sites was taken. 58 Web sites were randomly selected which had a total of 798,306 sessions (3.07%) (13,872 goal (20.50%)).

Step 3. Remove Single Page Sessions

The third step followed Bucklin and Sismeiro (2003) and removed all sessions which consisted of only a single page-view. A single page-view does not represent "browsing" behavior on a Web site (Bucklin and Sismeiro, 2003) and thus is unlikely to provide interesting visitor patterns. 181,699 single-page sessions were removed.

Step 4. Identify User-sessions

The fourth step of the process identified user-sessions from the remaining 616,607 sessions. A *user-session* encapsulates the session being analyzed (i.e., the "target") and all other sessions that met two requirements.

(1) The other session must have taken place at an e-commerce Web site.

(2) The other session must have ended between 30 minutes before the start of the target session and the end of the target session.

A valid user-session was needed to calculate relative measures for the user-centric clickstream model of information foraging. To be considered valid, a user-session must have had at least one session in addition to the target session[4].

Each of the 616,607 sessions from the third step was analyzed to determine if they were part of a valid user-session[5]. A total of 511,397 valid user-sessions (82.94%) were found and retained. The remaining 105,210 sessions (17.06%) did not have a valid user-session and were removed.

Step 5. Remove Sites with $<$ 50 Goal User-sessions

For the fifth step of the process, Web sites without at least 50 goal user-sessions were removed. Although step two initially checked for Web sites having at least 50 goal sessions, the number of goal user-sessions may have been reduced because a goal session may not have represented a valid user-session if no "other" sessions were associated with the target session.

A total of six Web sites (10.34%) were removed from the dataset because they had fewer than 50 goal user-sessions. The 9,266 user-sessions (1.81%) at those six Web sites were also removed.

[4]Specifying at least two sessions for a valid user-session is similar to requiring at least two page views for a valid session (Bucklin and Sismeiro, 2003).

[5]Further detail about how user-sessions were determined can be found in §5.1.1.

Step 6. Remove Outliers

The dataset was then examined for outliers in the sixth step of the process. An outlier was defined as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett and Lewis, 1994, pg. 7). Since the clickstream model of information foraging relies on relative behavior to other sessions, inconsistent sessions were removed from the dataset. Consistency was compared via the combination of the total number of pages viewed and the duration of a session.

An unsupervised density-based clustering algorithm called DBSCAN[6] (Ester et al., 1996) was used to locate outlying sessions[7]. DBSCAN identifies clusters of arbitrary shape, where the number of clusters is automatically determined via the algorithm. A cluster is formed by having a minimum number of neighbor points[8] ($MinPts$), or density, within a specified radius (Eps). Points not classified to a cluster are labeled as "noise" (i.e., outliers).

DBSCAN requires two user-specified parameters: $MinPts$ and Eps.

MinPts – the minimum number of points within a neighborhood of radius Eps. For two-dimensional datasets, $MinPts$ is commonly set to four (Ester et al., 1996; Hodge and Austin, 2004).

Eps – the Eps-neighborhood or radius of a cluster. The value of Eps is determined visually via a sorted k-dist graph (see point three below) (Ester et al., 1996).

To perform the outlier analysis using DBSCAN, four steps were followed.

(1) Goal and non-goal sessions were separated into two separate datasets. Each of the remaining three steps was performed independently on each dataset.

   (a) For the user-centric model the goal and non-goal datasets also included the "other" sessions associated with a user session. All other sessions that also achieved a goal were included in the goal dataset, while the remaining sessions were placed in the non-goal dataset.

[6]The average runtime complexity of DBSCAN is $O(n * log(n))$ (Ester et al., 1996).

[7]DBSCAN was chosen over common statistical techniques for removing outliers, such as removing values greater than three standard deviations away, for two reasons: (1) DBSCAN does not require knowledge of an underlying distribution and (2) DBSCAN is capable of finding outliers in multiple dimensions.

[8]The term points will be used to refer to sessions with a unique combination of pages viewed and session duration during the remainder of this subsection.

The goal dataset consisted of 20,121 goal sessions[9]. 10,834 of those sessions (53.84%) were target sessions, while the other 9,287 (46.16%) were other sessions. The non-goal dataset consisted of 1,589,407 non-goal sessions[10]. 491,297 of those sessions (30.91%) were target sessions, while the other 1,098,110 (69.09%) were other sessions.

(2) Values from each dimension were normalized between 0 and 1 according to equation 6.1, where x is a set of distinct values for a dimension, x_i is the i^{th} element of the set, and $min(x)$ and $max(x)$ are the minimum and maximum values found in set x, respectively.

$$norm(x_i) = \frac{x_i - min(x)}{max(x) - min(x)} \tag{6.1}$$

Normalization was done because distance calculations were used for generating both the sorted k-dist graph and for determining which points belonged within the same neighborhood. The Euclidean distance (equation 6.2) was used to calculate the distance between two points P and Q with n dimensions, such that $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$.

$$Euclidean\ distance = \sqrt{\sum_{i=0}^{n-1} (p_i - q_i)^2} \tag{6.2}$$

If values were not normalized, then distances may differ or not differ simply due to the scale of one dimension. For example, figure 34 illustrates the positions of three points: A, B, and C. Table 32 displays the non-normalized and normalized values for each of the point's two dimensions: number of pages viewed and session duration. Table 33 lists the Euclidean distance between pairs of points using each point's non-normalized and normalized values for both dimensions.

Using the non-normalized values from table 32, the distance (as seen in table 33) from A to C (45.00) is the same as the distance from B to C (45.00). However, looking at figure 34 it is apparent that an increase of 45 pages viewed represents a larger change in distance than a decrease of 45 minutes in session duration. The normalized distances for A to C (0.12) and B to C (0.45) better reflect the actual distance between points.

[9] A goal session may be present in more than one user-session.
[10] A non-goal session may be present in more than one user-session.

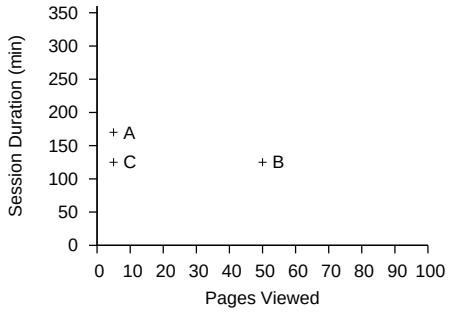


Figure 34.: Example Outlier Points

Table 32: Example Outlier Points

	Non-normalized		Normalized	
Point	Pages Viewed	Session Duration (min)	Pages Viewed[a]	Session Duration (min)[b]
A	5	170	0.05	0.47
B	50	125	0.50	0.35
C	5	125	0.05	0.35

[a] Assumes minimum and maximum values of 0 and 100 for pages viewed, respectively.
[b] Assumes minimum and maximum values of 0 and 360 for session duration, respectively.

Table 33: Example Outlier Distances

Points	Non-normalized Distance	Normalized Distance
A to B	63.64	0.47
A to C	45.00	0.12
B to C	45.00	0.45

(a) A random sample of 15.00% of the non-goal dataset's normalized points were selected for processing by DBSCAN. A sample was used because the time required to compute clusters from the original dataset using DBSCAN would have been prohibitively high[11]. The sample was used to find the boundary between non-outlying and outlying points. Points not included in the random sample that fell outside the non-outlying region were classified as outliers.

The entire non-goal dataset consisted of 1,589,407 non-goal sessions. A random sample of 238,411 sessions (15.00%) was selected and used in the remaining two steps.

(3) The parameters for DBSCAN were set to define the "thinnest" cluster in the dataset by following a three-step heuristic outlined by Ester et al. (1996). The "thinnest" cluster is the smallest or least dense grouping of points that are not considered noise.

(a) $MinPts$ was set to four since each dataset only had two dimensions (Ester et al., 1996)[12].

(b) The threshold distance, which distinguishes between noise and clusterable points, was located. Points farther away than the threshold distance (i.e., to the left) were considered "noise", while points closer than the threshold distance (i.e., to the right) were clusterable. To determine the threshold, a sorted k-dist graph was created ($k = MinPts$), where the distance of each point to its k^{th} neighbor is found, sorted in descending order, and then graphed. The purpose of the sorted-k-dist graph was to visually locate the first "valley" in distance values, which represents the threshold distance.

The Approximate Nearest Neighbor (ANN) search library version 1.1.1 (Mount and Arya, 2006) was used to calculate the distance of each point to its fourth nearest neighbor[13]. Figures 35a and 35b show the sorted 4-dist graphs of the first 100 values for the goal and non-goal sessions. Each figure was manually inspected to find the first "valley", which is shown at the intersection of dashed lines.

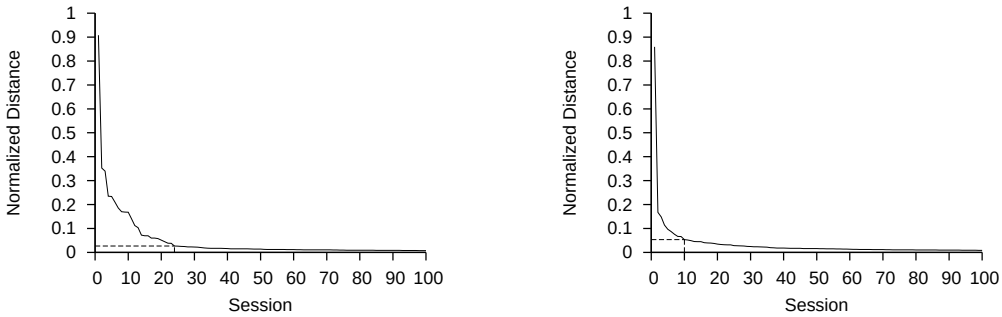
(c) Eps was set to the threshold distance found in step (b).

Table 34 lists the parameter values used for each of the datasets. $MinPts$ was set to four

[11] All points from the goal dataset were used.

[12] In a survey of outlier detection methodologies, Hodge and Austin (2004) also stated $MinPts$ is commonly set to four for DBSCAN.

[13] ANN is available at http://www.cs.umd.edu/~mount/ANN/.



(a) Goal Sessions (b) Non-goal Sessions

Figure 35.: User-centric: Sorted 4-Dist Graphs: Goal and Non-Goal

according to Ester et al. (1996) and Eps was determined from visually examining the k-dist graph for each dataset (figures 35a and 35b).

Table 34: User-centric: Parameter Values for DBSCAN

Sessions	MinPts	Eps
Goal	4	0.0266
Non-goal	4	0.0536

(4) The DBSCAN algorithm was run using RapidMiner Community Edition version 4.4[14] with the specified parameter values from table 34.

DBSCAN labeled 22 goal sessions (0.11%) as noise (i.e., outliers). Of the 22 goal outliers, four (18.18%) were from target sessions and the other 18 (81.82%) were from other sessions[15]. The non-outlying points all had durations of less than 500 minutes (8.33 hours) and viewed fewer than 800 pages.

A total of 7 non-goal outliers ($< 0.01\%$) were also found by DBSCAN in the random sample[16]. None of the outliers were from target sessions. The outliers found in the random sample were going to be used as boundary points to classify sessions from the entire non-goal dataset. However,

[14]RapidMiner is available at http://www.rapidminer.com. RapidMiner was previously named YALE (Yet Another Learning Environment).

[15]The 22 goal outlier sessions were represented by 22 distinct combinations of points.

[16]The 7 non-goal outlier sessions were represented by 7 distinct combinations of points.

after examining the results of the DBSCAN run, a number of sessions not flagged as outliers had extremely high values for either session duration or number of pages viewed. For example, a session in the random sample with a duration of 1,980 minutes (33 hours) was not flagged as an outlier, nor was a session with 10,000 pages viewed[17].

Although there were not a large number of sessions with extreme values, there were enough points within the same area to be considered a neighborhood by DBSCAN. In addition, there were enough of these small groups that were within a short distance of one another that they chained together to become part of the non-outlying cluster. Due to the difficulty in finding a clear separation between outlying and non-outlying points in the non-goal dataset, the boundaries found in the goal dataset were used for the non-goal dataset.

Using the cutoff values from the goal dataset ($\geq$ 800 pages viewed or $\geq$ an 800 minute session duration), 13,799 non-goal sessions (0.87%) were labeled as outliers. Of the 13,799 non-goal outliers, 89 (0.65%) were from target sessions and the other 13,710 (99.36%) were from other sessions[18].

Figures 36a and 36b show plots of the *distinct* outlier and non-outliers points for both the goal and non-goal sessions, respectively. Since only *distinct* points are shown in the figures, an accurate representation of the density of points in an area is difficult to determine.

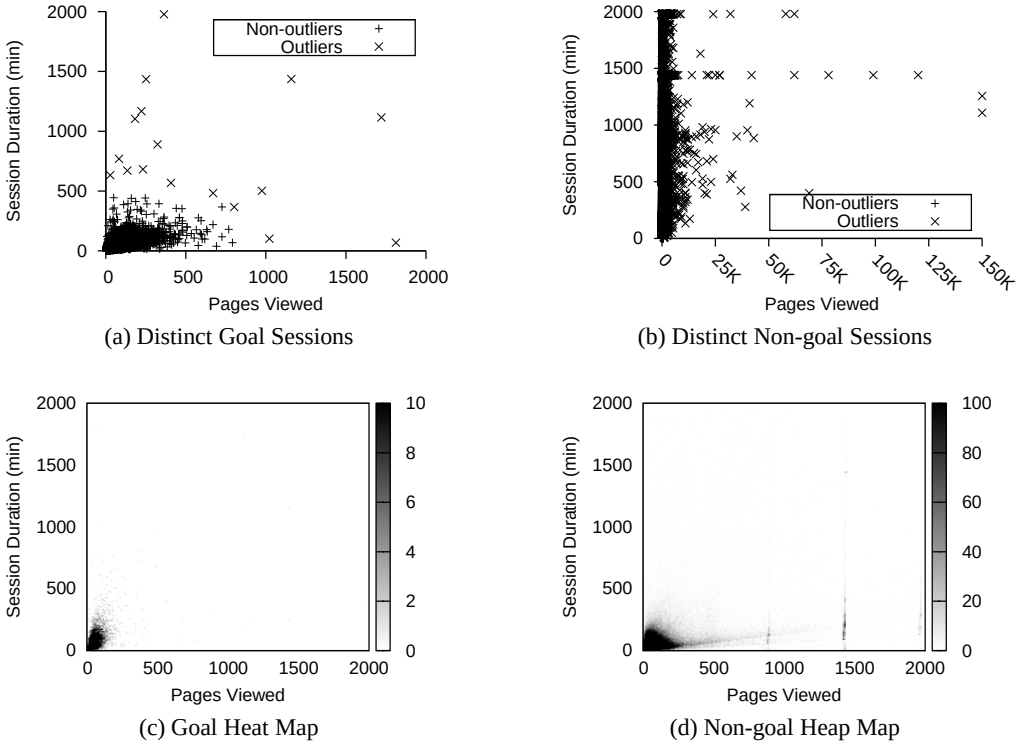
To illustrate density within an area, figures 36c and 36d present heat maps for the goal and non-goal datasets, respectively. The darkness in shade of each point in the heat map illustrates how many other sessions exist within the same area. A black point represents 10 or more sessions in the goal dataset, while 100 or more sessions are represented by the same shade in the non-goal dataset.

Noticeable within the non-goal heat map (figure 36d) is that even though figure 36b shows sessions with durations close to 2,000 minutes, the heat map demonstrates the density of points in those areas is practically non-existent. In addition, the figure also illustrates the use of the goal boundaries on the non-goal dataset retained the densest area of points as non-outlying sessions.

After removing all outliers, a total of 5,788 user-sessions (1.15%) were removed from the dataset.

[17] One possible explanation for sessions with extremely high values may be due to automated programs browsing the Web. For example, a program which resides as a background process may make a connection to a Web site to refresh its local cache of information every few minutes. If a user has an always-on Internet connection and does not turn off their computer, then it is feasible a session may last many hours. A similar argument can also be made for spidering programs that visit a large number of pages at a Web site.

[18] The 13,799 goal outlier sessions were represented by 12,642 distinct combinations of points (91.62%).



(a) Distinct Goal Sessions

(b) Distinct Non-goal Sessions

(c) Goal Heat Map

(d) Non-goal Heap Map

Figure 36.: User-centric: Outlier Points Plot

93 of those 5,788 user-sessions (1.61%) were removed because the target session was classified as an outlier. The remaining 5,695 user-sessions (98.39%) were removed because there was not at least one other session within the user-session (i.e., all the other sessions were classified as outliers and removed). A total of 496,343 user-sessions (10,714 goals) remained after processing all outliers.

Step 7. Remove Sites with $<$ 50 Goal User-sessions

For the final step of the process, Web sites without at least 50 goal user-sessions were removed. Although step five also checked for Web sites having at least 50 goal user-sessions, the number of goal user-sessions may have been further reduced due to the outlier analysis. If a user-session's target session, all "other" sessions, or both were flagged as outliers, then the user-session would become invalid.

No Web sites were removed, as all sites retained at least 50 goal user-sessions.

6.1.2 Final Dataset

The following subsections provide general statistics about the final dataset, along with characteristics of the Web sites and user-sessions in the dataset.

General Statistics

Table 35 displays general statistics for the final dataset. The first row of the table lists the total number of sessions in the dataset[19]. Each row after the first lists the total count for the metric and also its percentage compared to the total number of sessions. The overall conversion rate of the dataset was 2.16%, which is similar to the two percent conversion rate typically found at e-commerce Web sites (Moe, 2003; Sismeiro and Bucklin, 2004). Unique visitors accounted for 11.12% of the sessions whereas 88.88% of the sessions were from repeat visitors. Lastly, 7,366,442 pages were viewed over all 496,343 sessions from the 52 Web sites in the dataset.

[19]Unless otherwise specified all statistics are about the target session of each user-session. The term "session" will be used in place of "target session" for readability purposes.

Table 35: User-centric: Final Dataset Statistics

	n	%
Sessions	496,343	n/a
Goal sessions	10,714	2.16%
Non-goal sessions	485,629	97.84%
Unique visitors	55,195	11.12%
Repeat visits	441,148	88.88%
Pages viewed	7,366,442	n/a
Web sites	52	n/a

Web Site Characteristics

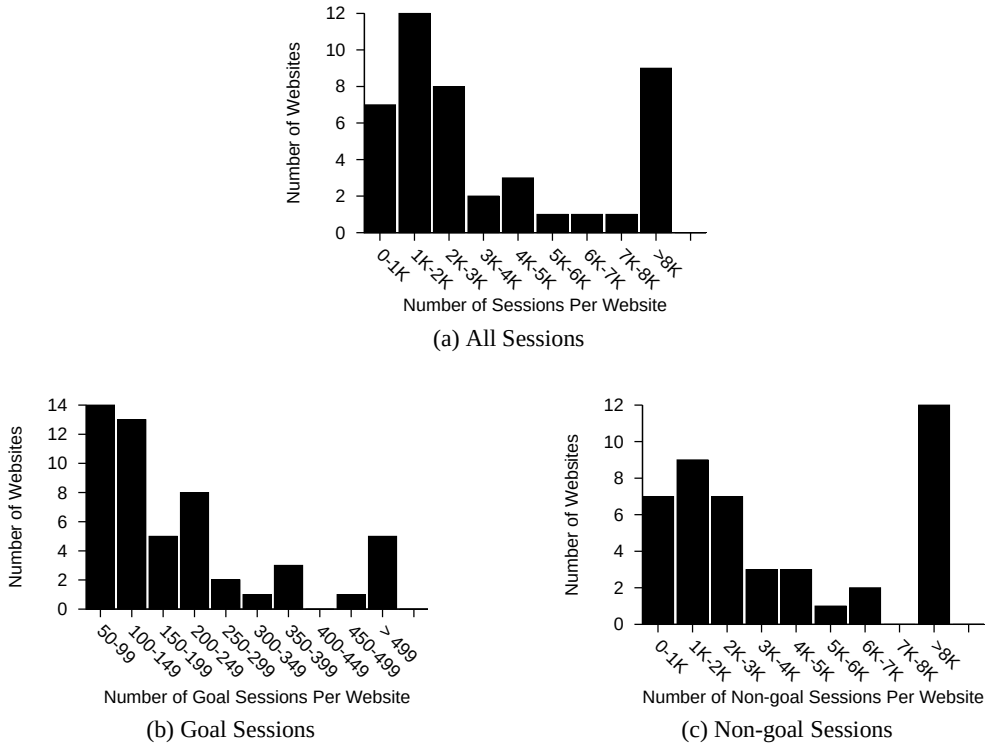
Table 36 provides the mean, standard deviation, minimum, and maximum values from all 52 Web sites for the number of goal, non-goal, and total sessions visiting each site and the conversion rate from each site.

Table 36: User-centric: Web Site Characteristic Statistics

	Mean	St. Dev.	Minimum	Maximum
SESSIONS				
Total sessions	9,545.06	14,078.48	406	76,138
Goal sessions	206.04	154.91	51	597
Non-goal sessions	9,339.02	14,064.45	290	76,041
OTHER				
Conversion	6.70%	7.17%	0.12%	28.57%

On average, each Web site had 9,545.06 sessions visiting the Web site, with more than 45 times as many non-goal sessions as goal sessions. Each Web site had, on average, 206.04 goal sessions (2.16%) and 9,339.02 non-goal sessions (97.84%).

Figures 37a – 37c illustrate the distribution of the number of total, goal, and non-goal sessions for each Web site, respectively. The majority of Web sites (32 out of 52 (61.54%)) had fairly light traffic, having less than 8,000 total sessions (figure 37a). However, there were 20 Web sites (38.46%) with more than 8,000 total sessions, with the most heavily-visited site having 76,138 total sessions. In terms of goal sessions (figure 37b), 40 Web sites (76.92%) had between 50 and 249 goal sessions, with the remaining 12 Web sites (23.08%) having more than 250 goal sessions.



(a) All Sessions

(b) Goal Sessions

(c) Non-goal Sessions

Figure 37.: User-centric: Web site Sessions Histograms

The average conversion rate for the 52 Web sites was 6.70%, with one Web site having the highest rate of 28.57%. Figure 38 illustrates the distribution of conversion rates for each Web site. 22 of the 52 Web sites (42.31%) had less than a 3% conversion rate. 15 of the Web sites (28.85%) had between a 3% and 8% conversion rate. The remaining 15 Web sites (28.85%) had a conversion rate higher than 8%.

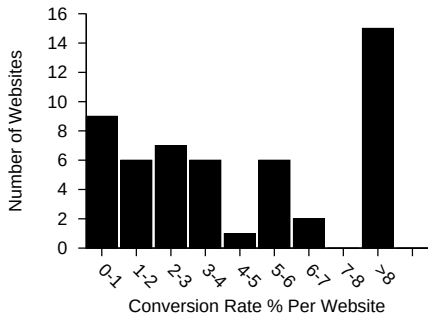


Figure 38.: User-centric: Web site Conversion Histogram

Session Characteristics

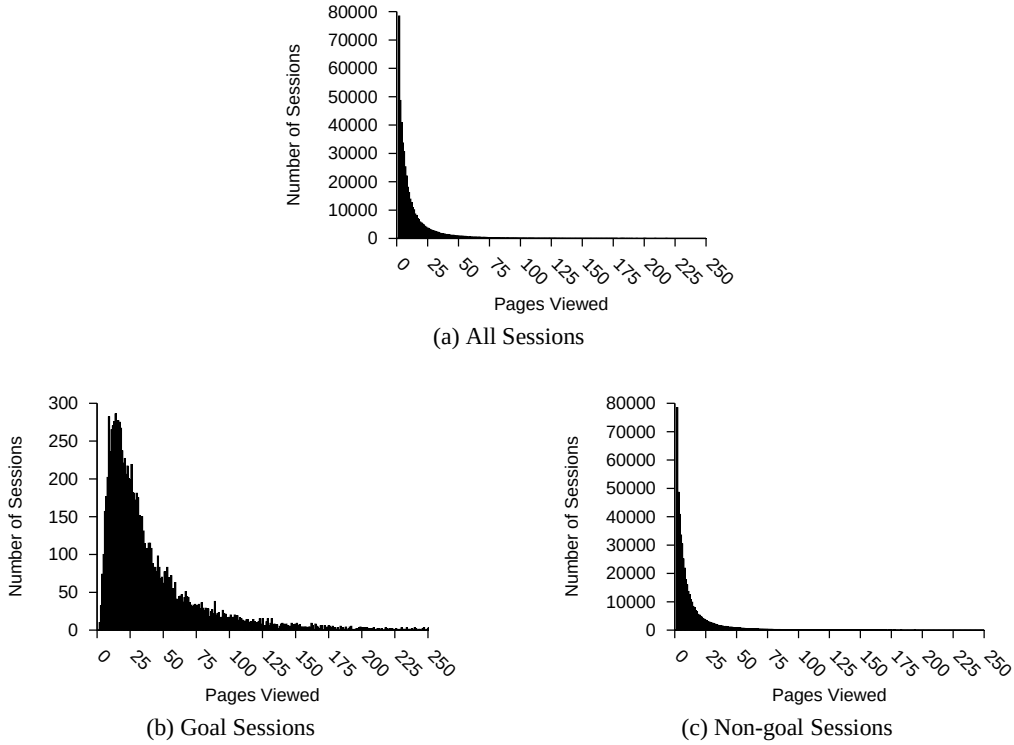
Table 37 provides the mean, standard deviation, minimum, and maximum values for the number of pages viewed and duration from all 496,343 sessions in the dataset. For each metric, values are provided for three sets of sessions: goal, non-goal, and all sessions.

Table 37: User-centric: Session Characteristic Statistics

	Mean	St. Dev.	Minimum	Maximum
PAGES VIEWED				
All sessions	14.84	26.07	2	794
Goal sessions	42.33	48.70	2	791
Non-goal sessions	14.24	25.01	2	794
SESSION DURATION (MIN)				
All sessions	10.17	16.88	1	495
Goal sessions	27.31	26.56	1	367
Non-goal sessions	9.79	16.40	1	495

Each session consisted, on average, of less than 15 page views (14.84), with a maximum of 794 pages viewed by one session. Goal sessions viewed almost three times as many pages per session, on average, compared to non-goal sessions (42.33 versus 14.24). Figures 39a – 39c show the distribution of pages viewed by number of sessions.

The average duration from all 496,343 sessions was 10.17 minutes, with one session spending over 495 minutes (8.25 hours) on a site. Goal sessions spent almost three times as many minutes



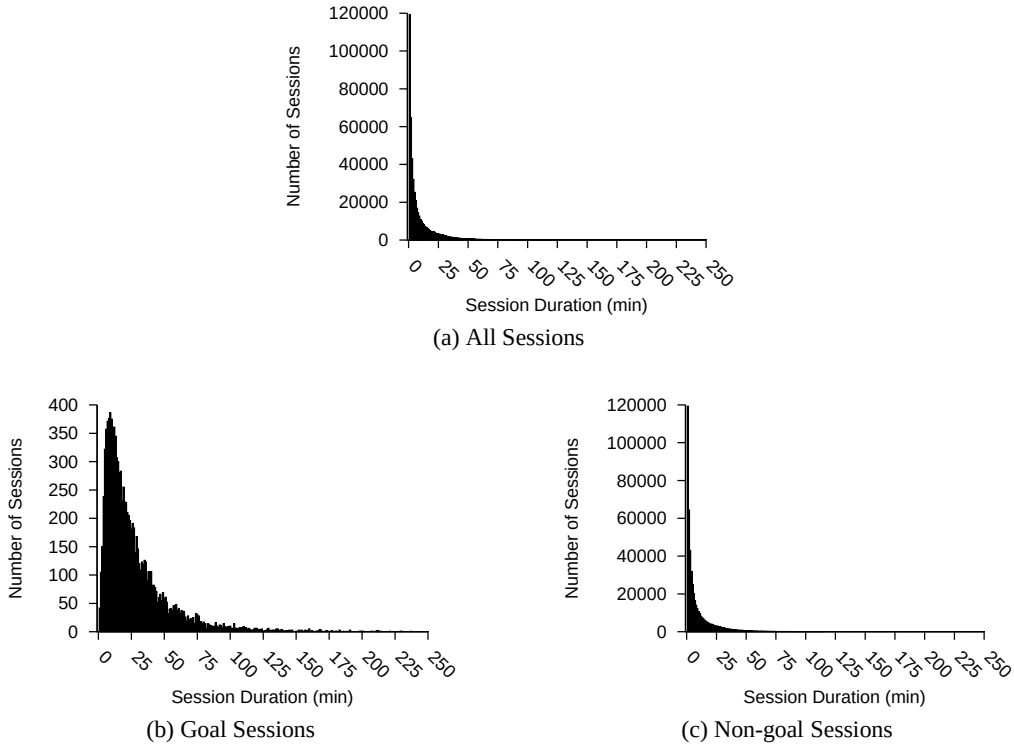
(a) All Sessions

(b) Goal Sessions

(c) Non-goal Sessions

Figure 39.: User-centric: Session Pages Viewed Histograms

on a site compared to non-goal sessions (27.31 min versus 9.79 min). Figures 40a – 40c illustrate the distribution of session duration in minutes by number of sessions.



(a) All Sessions

(b) Goal Sessions

(c) Non-goal Sessions

Figure 40.: User-centric: Session Duration Histograms

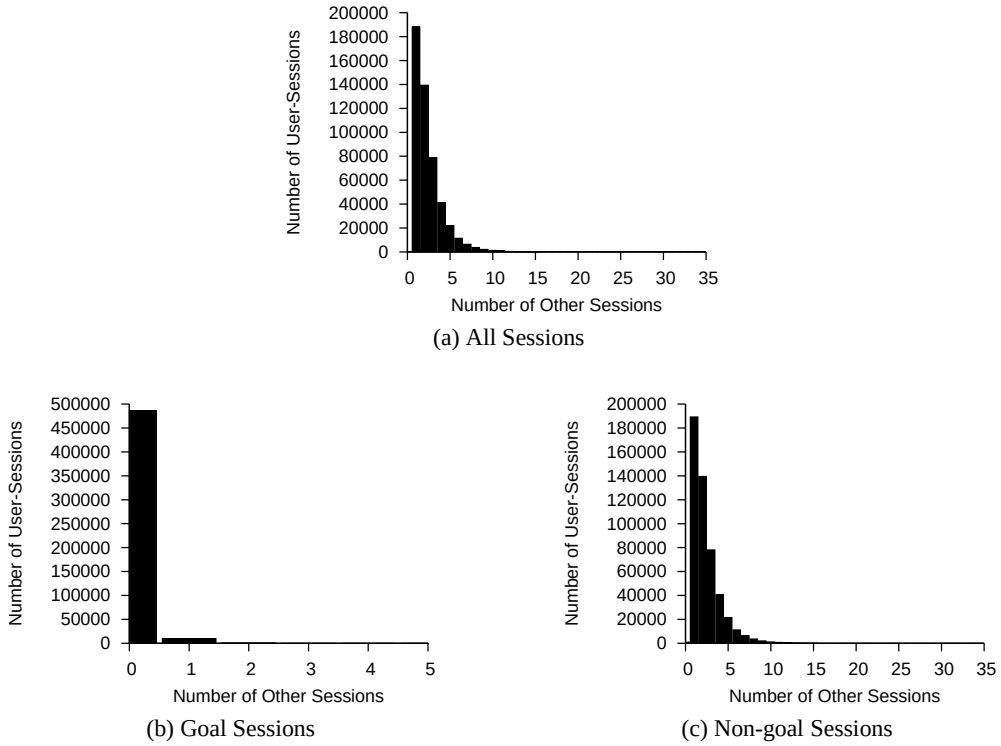
User-session Characteristics

Table 38 provides the mean, standard deviation, minimum, and maximum values for the number of total, goal, and non-goal other sessions for all 496,343 user-sessions in the dataset.

Table 38: User-centric: User-session Characteristic Statistics

	Mean	St. Dev.	Minimum	Maximum
NUMBER OF OTHER SESSIONS				
All sessions	2.37	1.70	1	32
Goal sessions	0.02	0.15	0	4
Non-goal sessions	2.35	1.69	0	32

Each user-session consisted, on average, of less than three other sessions (2.37), with a maximum of 32 other sessions by one user-session. Other sessions were mostly comprised of non-goal sessions (2.35 versus 0.02). However, one user-session had four other sessions that were goals. Figures 41a – 41c show the distribution of other sessions by number of user-sessions.



(a) All Sessions

(b) Goal Sessions

(c) Non-goal Sessions

Figure 41.: User-centric: User-session Sessions Histograms

6.2 Site-centric Dataset

The data used for the site-centric clickstream model of information foraging was provided by a Web hosting company. The data was captured over a year period from September 12, 2007 to September 23, 2008. The web hosting company was unique since it provided a common platform for Web sites of a similar nature. For example, their Web sites all used the same platform that allowed the site owners to add content to their Web site without knowledge of HTML. A beneficial byproduct of having sites on the same platform was a common structure to each Web site. For example, those Web sites with a contact form all submitted their contact information to the same

common platform URL. The contents of the contact form were then saved and a result page was displayed to the user[20]. Therefore, it could be determined that a goal was achieved (i.e., a visitor filled out a contact form and submitted it) if a visitor's session "viewed" the contact form submission page.

Since the data provider hosted thousands of Web sites, they created a mechanism to capture traffic from all their sites without relying on the individual traffic logs of each Web site. Whenever a user visited a Web page of a participating Web site a small transparent image was downloaded via a JavaScript script. The image had parameters unique to the user along with information such as the Web site and Web page being visited, timestamp of visit, and other miscellaneous information. Once the script was deployed on the platform, it was integrated into Web sites once site owners updated their site in some way (e.g., a page was edited). Therefore, even though the script was deployed on September 12, 2007, data collection at a particular Web site only started once the site was changed in some way.

Each piece of data was stored in a data warehouse and linked to the user, Web site, and Web page it referenced. A visitor's session was defined as any sequence of Web pages on the same Web site by the same visitor with less than a 30 minute time period between page viewings. A 30 minute session timeout has also been used in previous clickstream research (Bucklin and Sismeiro, 2003; Sismeiro and Bucklin, 2004; Van den Poel and Buckinx, 2005).

The remainder of this section provides information regarding the steps taken to arrive at the final dataset, along with general descriptive statistics about the data. The preprocessing steps applied to the data are described in the next section. Descriptive statistics are then provided in the following section.

6.2.1 Preprocessing of Original Dataset

The data obtained from the data provider included many data elements not applicable to the current research. Therefore, a number of processing steps were performed to obtain a final dataset usable for testing the site-centric clickstream model of information foraging. Table 39 lists each step of the process along with the total number of Web sites, sessions, and goal sessions; and how many Web sites, sessions, and goal sessions were removed at that step (if applicable). Table 40

[20]The result page may be a return to the contact form that was submitted, a page thanking the user for submitting their information, or any other page on the Web site (e.g., the index page).

lists the parameters used in each preprocessing step. A discussion of each step and its parameters are provided below.

Table 39: Site-centric: Preprocessing of Original Dataset Statistics

Step	Description	Web Sites	Sessions	Goals	△ in Web Sites	△ in Sessions	△ in Goals
	Original dataset	6,003	1,968,491	n/a	n/a	n/a	n/a
1	Map valid pages	6,003	1,968,491	n/a	n/a	n/a	n/a
2	Remove other Web sites	1,710	1,692,275	n/a	−4,293	−276,216	n/a
3	Remove spam sessions	1,504	1,689,159	n/a	−206	−3,116	n/a
4	Remove single page sessions	1,483	900,677	n/a	−21	−788,482	n/a
5	Determine goal sessions	1,483	900,677	12,441	n/a	n/a	n/a
6	Remove no goal Web sites	918	790,691	12,441	−565	−109,986	n/a
7	Remove Web sites with < 50 goal sessions	57	278,463	5,982	−861	−512,228	−6,459
8	Remove outliers	57	278,437	5,975	n/a	−26	−7
9	Identify contact goals	57	278,437	5,975	n/a	n/a	n/a
10	Classify goal sessions	57	278,437	5,827	n/a	n/a	−148
11	Remove Web sites without any contact goals having ≥ 50 goal sessions	47	250,162	5,302	−10	−28,275	−525
12	Classify other contact goal sessions as non-goal sessions	47	250,162	4,979	n/a	n/a	−323

Table 40: Site-centric: Preprocessing Parameters

Step	Description	Parameters
	Original dataset	n/a
1	Map valid pages	n/a
2	Remove other Web sites	$platform \neq informational$
3	Remove spam sessions	$sessions == spam$
4	Remove single page sessions	$sessionLength == 1$
5	Determine goal sessions	$formSubmissionPage \neq visited$
6	Remove no goal Web sites	$goalsAtWebsite == 0$
7	Remove Web sites with < 50 goal sessions	$goalsAtWebsite < 50$
8	Remove outliers	$MinPts = 4$ $Eps = 0.0636$ (goal sessions) $Eps = 0.0597$ (non-goal sessions)
9	Identify contact goals	$countedSupport = 5$ $patternSize = 3$ $pattern = AXA$ or AXB $directMatches = 5$
10	Classify goal sessions	$0 \leq gap < sessionLength - 1$
11	Remove Web sites without any contact goals having ≥ 50 goal sessions	$goalsAtContactGoal < 50$
12	Classify other contact goal sessions as non-goal sessions	n/a

Step 1. Mapping Valid Web pages

The first step of the process removed "invalid pages" from the dataset and mapped "valid" pages together. Since the data provider relied on a JavaScript script to provide information on which page was visited, there were instances where the actual page visited could not be determined. For example, if a user visited http://www.domain.com/mypage.html then it would be recorded that domain.com/mypage.html was visited (i.e., a valid page). However, if the user viewed mypage.html through a service such as Google's cache, then the URL recorded for the user might be something like 30.186.56/search/cache. Since there was no way to determine what page was actually viewed

in Google's cache, such pages were eliminated from the dataset (i.e., an invalid page).

Instead of examining each page to determine if it was valid or invalid, the domains for each page were examined instead. First, any page from a domain which was present in more than one Web site was considered invalid[21]. Second, pages from many search engine caches are recorded with an IP address rather than a domain name. Therefore, any pages from a numerical IP address were considered invalid. Finally, a manual inspection of the remaining domains was done to remove known outside services (e.g., Web-based mail domains).

In addition to removing invalid pages, valid pages needed to be mapped together on the same Web site. Web sites with multiple domain names pointing to the same Web site would only show a fragmented picture of the pages being visited. For example, assume domainA.com and domainB.com both point to the same Web site. In the data, domainA.com/mypage.html and domainB.com/mypage.html would be seen as totally separate pages from one another. Instead, a visit to mypage.html should be counted as the same page, as long as the domain was valid. Thus, pages of the same name were mapped to a single valid page.

A total of 43,544 unique pages were present in the entire dataset. 5,702 of those pages (13.09%) were flagged as invalid. Of the remaining 37,842 unique pages, 4,102 of those (10.84%) were mapped to other existing pages (e.g., domain.com/ mapped to domain.com/index.html). After all processing was done, a total of 33,740 unique valid pages remained.

Step 2. Remove Other Web sites

After completing the first step, the dataset still retained the original 6,003 Web sites and 1,968,491 sessions. However, the dataset included data from other platforms the data provider hosted (e.g., social networking Web sites) which were not the focus of this research. Therefore, the second step of the process removed all Web sites not using the data provider's informational platform. A total of 4,293 Web sites were removed along with 276,216 corresponding sessions.

[21]The data provider offered a number of services that used the same domain on multiple Web sites. Those domains were flagged as "invalid" even though the origin of the domain was known. However, this did not affect the analysis since Web sites using those shared domains were from other platforms (e.g., social networking) and were not being investigated in this research.

Step 3. Remove Spam Sessions

The third step removed any sessions designated as spam. The data provider flagged any sessions from robots, spiders, or any other automated browsing mechanisms as spam (e.g., Google's indexing spider). A total of 3,116 sessions and 206 Web sites (which only had spam sessions) were removed.

Step 4. Remove Single-page Sessions

The fourth step followed Bucklin and Sismeiro (2003) and removed all sessions which consisted of only a single page-view[22]. A single page-view does not represent "browsing" behavior on a Web site (Bucklin and Sismeiro, 2003) and thus is unlikely to provide interesting visitor patterns. 788,482 single-page sessions were removed along with 21 Web sites which only had single page sessions.

Step 5. Determine Goal Sessions

For the fifth step of the process, sessions were classified as either "goal" or "non-goal" sessions. Within the data, any contact form submission was represented as a visit to a specific URL (e.g., formSubmission.html). A total of 12,441 sessions (1.38%) visited the Web site-unique form submission URL, and thus were classified as goal sessions.

1,857 of the goal sessions (14.93%) visited the form submission URL more than once during a session (i.e., a repeat goal session). 1,563 of the repeat goal sessions (84.17%) were instances where the form submission page was visited multiple times in a row. A potential explanation for this behavior is the user clicked the submit button on a form multiple times.

The remaining 294 repeat goal sessions (15.83%) submitted a form and then visited at least one other page before submitting a form again (i.e., distinct form submissions)[23]. Such repeat behavior may be the result of a user submitting different contact forms on a Web site (e.g., request for information and signing up for a newsletter), or may be a person simply resubmitted the same form (for whatever reason) after going somewhere else on the site. Figure 42 shows a histogram of the

[22] Only sessions with more than one *valid* page viewed were retained.

[23] The 294 sessions with distinct form submissions were retained in the analysis. Only browsing behavior occurring before the *first* form submission was considered in the analysis, and thus having extra form submissions did not impact the analysis for this research.

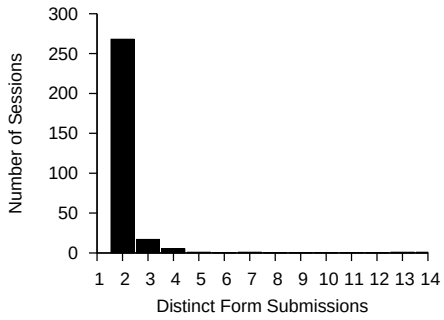


Figure 42.: Site-centric: Distinct Form Submissions Histogram

number of distinct form submissions by number of repeat goal sessions. The maximum number of distinct form submissions was 14.

Step 6. Remove No-goal Web sites

The sixth step removed Web sites which did not have any goals achieved during the data collection period. A total of 565 Web sites were removed along with the 109,986 sessions which occurred on those Web sites.

Step 7. Remove Web sites with Fewer Than 50 Goal Sessions

In order to ensure a large enough sample size of goal sessions for analysis, the seventh step removed Web sites which had fewer than 50 goal sessions. 861 Web sites were removed along with the 512,228 corresponding sessions at those sites.

Prior to the removal of Web sites in this step, a cutoff point was determined by examining a histogram of the number of Web sites according to the number of goal sessions at their site (figure 43). 98 Web sites (10.68%) with 30 goal sessions or more are displayed in the figure[24]. Of those 98 Web sites shown, 41 of them (41.84%) had fewer than 50 goal sessions. 31 of those 41 sites (75.61%) only had between 30 and 39 goal sessions. Thus, the selection of 50 goal sessions as a cutoff point appears to be a good selection for including the maximum number of Web sites while ensuring a large enough goal session sample size within each site for the analysis.

[24]To provide a reasonable scale for the y-axis, the figure does not show the 820 Web sites (89.32%) with fewer than 30 goal sessions.

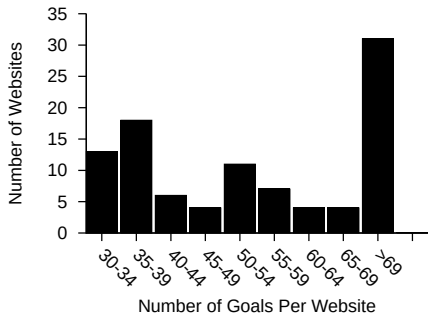


Figure 43.: Site-centric: Goal Sessions by Web site Histogram

Step 8. Remove Outliers

The dataset was then examined for outliers in the eighth step of the process. An outlier was defined as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett and Lewis, 1994, pg. 7). Inconsistent sessions were removed from the dataset. Consistency was compared via the combination of the total number of pages viewed and the duration of a session.

An unsupervised density-based clustering algorithm called DBSCAN[25] (Ester et al., 1996) was used to locate outlying sessions[26]. DBSCAN identifies clusters of arbitrary shape, where the number of clusters is automatically determined via the algorithm. A cluster is formed by having a minimum number of neighbor points[27] ($MinPts$), or density, within a specified radius (Eps). Points not classified to a cluster are labeled as "noise" (i.e., outliers).

DBSCAN requires two user-specified parameters: $MinPts$ and Eps.

MinPts – the minimum number of points within a neighborhood of radius Eps. For two-dimensional datasets, $MinPts$ is commonly set to four (Ester et al., 1996; Hodge and Austin, 2004).

Eps – the Eps-neighborhood or radius of a cluster. The value of Eps is determined visually via a sorted k-dist graph (see point three below) (Ester et al., 1996).

To perform the outlier analysis using DBSCAN, four steps were performed. The process is very

[25]The average runtime complexity of DBSCAN is $O(n * log(n))$ (Ester et al., 1996).

[26]DBSCAN was chosen over common statistical techniques for removing outliers, such as removing values greater than three standard deviations away, for two reasons: (1) DBSCAN does not require knowledge of an underlying distribution and (2) DBSCAN is capable of finding outliers in multiple dimensions.

[27]The term points will be used to refer to sessions with a unique combination of pages viewed and session duration during the remainder of this subsection.

similar to the user-centric process except user-sessions were not used for the site-centric dataset and thus "other" sessions were not included in the datasets. In addition, random sampling of the non-goal dataset was not used because of the smaller-sized site-centric dataset.

(1) Goal and non-goal sessions were separated into two separate datasets. Each of the remaining three steps was performed independently on each dataset.

(2) Values from each dimension were normalized between 0 and 1 according to equation 6.3, where x is a set of distinct values for a dimension, x_i is the i^{th} element of the set, and $min(x)$ and $max(x)$ are the minimum and maximum values found in set x, respectively.

$$norm(x_i) = \frac{x_i - min(x)}{max(x) - min(x)} \tag{6.3}$$

(3) The parameters for DBSCAN were set to define the "thinnest" cluster in the dataset by following a three-step heuristic outlined by Ester et al. (1996). The "thinnest" cluster is the smallest or least dense grouping of points that are not considered noise.

   (a) $MinPts$ was set to four since each dataset only had two dimensions (Ester et al., 1996)[28].

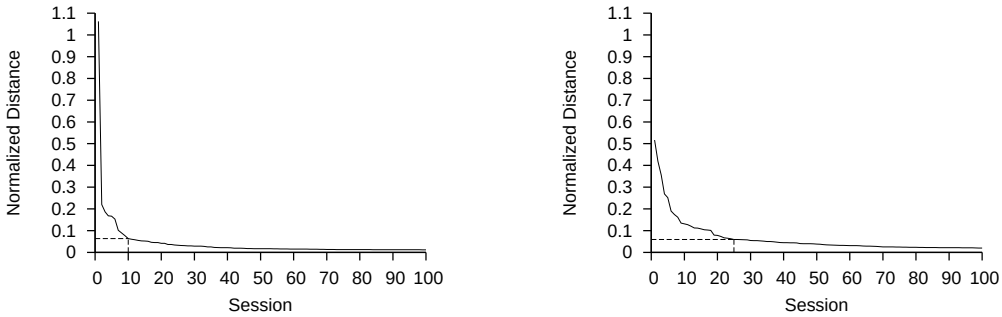
   (b) The threshold distance, which distinguishes between noise and clusterable points, was located. To determine the threshold, a sorted k-dist graph was created ($k = MinPts$), where the distance of each point to its k^{th} neighbor is found, sorted in descending order, and then graphed. The purpose of the sorted-k-dist graph was to visually locate the first "valley" in distance values, which represents the threshold distance.

   The Approximate Nearest Neighbor (ANN) search library version 1.1.1 (Mount and Arya, 2006) was used to calculate the distance of each point to its fourth nearest neighbor[29]. Figures 44a and 44b show the sorted 4-dist graphs of the first 100 values for the goal and non-goal sessions. Each figure was manually inspected to find the first "valley", which is shown at the intersection of dashed lines.

   (c) Set Eps to the threshold distance found in step (b).

[28] In a survey of outlier detection methodologies, Hodge and Austin (2004) also stated $MinPts$ is commonly set to four for DBSCAN.

[29] ANN is available at http://www.cs.umd.edu/~mount/ANN/.



(a) Goal Sessions (b) Non-goal Sessions

Figure 44.: Site-centric: Sorted 4-Dist Graphs: Goal and Non-Goal

Table 41 lists the parameter values used for each of the datasets. $MinPts$ was set to four according to Ester et al. (1996) and Eps was determined from visually examining the k-dist graph for each dataset (figures 44a and 44b).

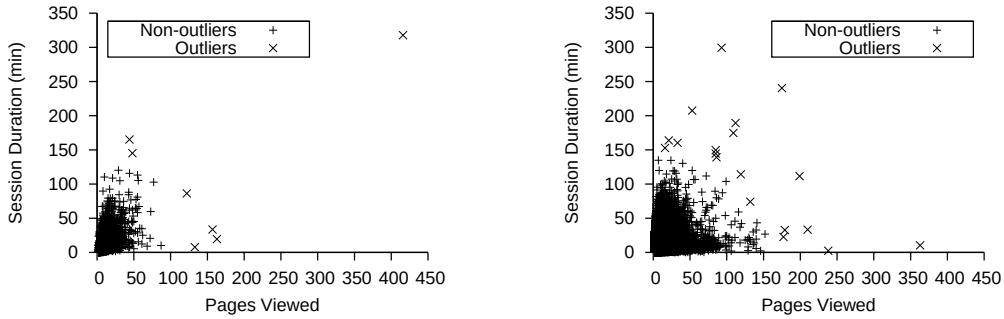
Table 41: Site-centric: Parameter Values for DB-SCAN

Sessions	MinPts	Eps
Goal	4	0.0636
Non-goal	4	0.0597

(4) The DBSCAN algorithm was run using RapidMiner Community Edition version 4.4[30] with the specified parameter values from table 41.

DBSCAN labeled seven goal sessions (0.12%) and 19 non-goal sessions (0.01%) as noise (i.e., outliers). Figures 45a and 45b show plots of distinct outlier and non-outliers points for both the goal and non-goal sessions, respectively. Roughly speaking, goal sessions with over 100 pages viewed or a duration of 145 minutes or more were considered outliers. For the non-goal sessions, there was not a clear separation between outliers and non-outliers for global values of pages viewed or session duration. Instead, figure 45b illustrates how different combinations of pages viewed and session duration categorized sessions as outliers or not.

[30]RapidMiner is available at http://www.rapidminer.com. RapidMiner was previously named YALE (Yet Another Learning Environment).



(a) Goal Sessions (b) Non-goal Sessions

Figure 45.: Site-centric: Outlier Points Plot

Step 9. Identify Contact Goals

For the ninth step of preprocessing, contact goals at each Web site were identified. A contact goal is the submission of a particular contact form on a Web site. A Web site may have more than one contact goal. For example, a Web site may have one contact form for general inquiries (contact goal A) and another contact form to request quotes (contact goal B). As a forager's information goal may differ drastically depending on the contact form being submitted, simply grouping all goal sessions together may introduce noise into the analysis. Classifying goal sessions by contact goal attempts to reduce noise by only grouping foragers together with similar information goals.

The eventual observable outcomes of this preprocessing step were three-fold:

(1) Identify contact goals having at least 50 goal sessions. The selection of 50 goals sessions was made to balance the need for sufficient sample size of goal sessions within a single contact goal and to include as many Web sites as possible. The actual selection of a single contact goal for a Web site is discussed in preprocessing step 12.

(2) Identify pages which were necessary conditions for the submission of a contact goal (e.g., contact form page, thank you page). Once identified, these necessary condition pages were then excluded from future mining of patches and trails.

(3) Classify goal sessions to an identified contact goal (discussed further in preprocessing step 10)[31].

[31] Not all goal sessions would be classifiable to a contact goal. This preprocessing process was only concerned with discovering moderately-visited contact goals. Thus, goal sessions which submitted forms for non-discovered contact goals would not be classified.

Within the data, any contact form submission was represented as a visit to a specific URL (e.g., formSubmission.html). Therefore, a forager submitting from either contact goal A or contact goal B would both show a visit to page formSubmission.html within their session. The limitation of this approach is not being able to directly classify goal sessions according to their contact goal. Therefore, an indirect manner of discovering contact goals via browsing patterns was used.

The general pattern of a form submission consisted of three pages in sequence: (1) a contact form page, (2) a page representing a form submission, and (3) a thank you page or the same contact form page from (1). To discover these sequences, frequent sequential patterns were mined using the Sequential Pattern Mining (SPAM) algorithm (Ayres et al., 2002)[32].

Potential Contact Goal To be considered a potential contact goal, a mined sequential pattern must have met five criteria.

(1) Have a counted support of at least five goal sessions. Although the interest in this processing was on contact goals with at least 50 goal sessions, the counted support was set to five for two reasons.

   (a) The first reason was to account for valid, but non-standard browsing behavior. For example, although the general submission pattern consists of three pages, there are occasions where a forager will only complete the first two pages of the sequence. This is because after the form submission, the system automatically forwarded a forager (after a short delay) to the third page. However, due to the delay some foragers may browse elsewhere or leave the site before being automatically forwarded.

   (b) The second reason was to discover as many contact goals as possible so that goal sessions were not incorrectly classified to the wrong contact goal. The browsing patterns of a forager may match, to differing degrees, multiple contact goals. If only highly-visited contact goals were discovered, a session may be classified to that contact goal

[32] Another method of discovering contact goals would be to use the referrer field of the form submission page to discover all contact form pages. However, the data provider limited the referrer field in this dataset to the domain-level. In addition, after submitting a form, a forager is automatically forwarded to the third page. This forwarding is done server-side and thus the referrer field would not be populated. Therefore, the pages shown as a result of a submitted contact form (e.g., a thank you page) could not be discovered by searching for the URL of the form submission page in the third page's referrer field. Due to these data and mechanism limitations, sequential pattern mining was used.

even though a less-visited contact goal was a better match. Classifying sessions to contact goals is discussed further in preprocessing step 11.

(2) Have a three-page sequence length.

(3) The first page of the sequence must *not* have been an index page or a form submission page.

(4) The second page of the sequence *must* have been a form submission page.

(5) The third page of the sequence must *not* have been a form submission page.

Confirmed Contact Goal A potential contact goal becomes a confirmed contact goal if it met the requirements listed above plus one additional requirement.

(1) A minimum of five goal sessions must directly match the pattern. A direct match means a goal session visited the exact same three pages, in order, and without any additional pages in between any of the pages of the sequence. The selection of a value less than 50 was again due to valid, but non-standard browsing behavior. For example, assume a contact goal consists of the pattern: pageA pageF pageA. A forager may visit the contact form page (pageA), open a new tab for the index page (pageI), and then return to the first tab and submit the contact form page. The session of the forager would be recorded as pageA pageI pageF pageA. Even though this session does not exactly match the pattern for the contact goal, it would still be considered a submission for the contact goal.

Conflicting Contact Goals During the process of discovering contact goals, four Web sites were flagged as having conflicting contact goals. A conflicting contact goal is where the same page either before or after the form submission is shared by another contact goal on the same Web site. For example, a conflict would occur if two contact goals on the same Web site have the same third page (e.g., contact.html) but different first pages (e.g., contact.html and product.html).

Table 42 provides information about the four Web sites and their conflicting contact goals. The first three Web sites (A-C) each had two conflicting contact goals while the fourth Web site (D) had three conflicts. For each conflicted contact goal the table lists the contact goal id, sequential pattern for the contact goal, and the number of direct sessions matched to the contact goal. The final column of the table describes the action taken to resolve the conflict for the Web site.

Table 42: Site-centric: Conflicting Contact Goals

Web site	Contact Goal	Page Pattern	Direct Matches	Action
A	CG-1	1. contactus.html 2. submission.html 3. contactus.html	267	Remove CG-2
	CG-2	1. productABC.html 2. submission.html 3. contactus.html	13	
B	CG-1	1. contactus.html 2. submission.html 3. contactus.html	104	Remove CG-2
	CG-2	1. products.html 2. submission.html 3. contactus.html	5	
C	CG-1	1. contactus.html 2. submission.html 3. contactus.html	550	Remove CG-2
	CG-2	1. productABC.html 2. submission.html 3. contactus.html	7	
D	CG-1	1. signup.html 2. submission.html 3. thanks.html	70	Combine all
	CG-2	1. signup1.html 2. submission.html 3. thanks.html	99	
	CG-3	1. signup1.html 2. submission.html 3. signup1.html	88	

The conflict between the contact goals on the first three Web sites shared three common characteristics.

(1) A highly-visited contact goal with a symmetrical page pattern (e.g., contactus.html, submission.html, and then contactus.html again).

(2) A rarely-visited contact goal with an asymmetrical page pattern (e.g., products.html, submission.html, and then contactus.html).

(3) The third page of the sequential pattern was shared between both contact goals.

To resolve the conflict between the contact goals at the first three Web sites, the highly-visited contact goals were retained and the rarely-visited contact goals were flagged as "invalid" and removed. The decision to remove the rarely-visited contact goals was made for two reasons.

(1) Symmetric patterns were the most common pattern found amongst contact goals. This is because the default behavior for Web sites on the informational platform was to return a user back to the original contact form after a form was submitted. Therefore, if one contact goal is symmetric, the other is asymmetric, and they both share the same third page, it is more likely the symmetric contact goal is valid.

(2) The rarely-visited contact goals likely represent indirect matches of the highly-visited contact goal. In other words, the sessions with a direct match to the rarely-visited contact goal were really indirect matches to the highly-visited contact goal. However, enough sessions visited the same page after the contact page, but before submitting the contact form (e.g., contactus.html, products.html, submission.html, and then contactus.html), to be discovered as a contact goal. This rationale is plausible because (1) there were so few direct matches for each rarely-visited contact goal and (2) of the 25 direct matches for rarely-visited contact goals, 24 of them (96.00%) were indirect matches for the highly-visited contact goal[33].

For the final Web site (D), none of the conflicting contact goals fully met the two reasons listed above to be considered "invalid" contact goals. In regards to the first point listed, although CG-1 and CG-2 shared the same third page (thanks.html), neither of the contact goals had a symmetric

[33]The single non-match did not visit any other pages except for the pattern for Web site A contact goal CG-2.

pattern. In addition, unlike the second point mentioned, all three contact goals had a substantial number of direct matches and no indirect matches were found between CG-1 and CG-2[34]. Therefore, the conflicting contact goals were examined to determine if they represented the evolution of a single contact goal's structure (e.g., changing names of forms, thank you behavior).

Informally, the site was hypothesized to contain a single contact goal (CG-A) for signing people up for activities that evolved from CG-1 to CG-2, and then to CG-3. Table 43 illustrates the first and third pages used for each contact goal (columns two through four). Initially, CG-A was believed to contain the pages signup.html and thanks.html (CG-1). However, at some point the page signup.html was replaced or renamed on CG-A with signup1.html (CG-2). Later on, CG-A was changed a third time when the thank you page (thanks.html) was dropped and the first page was also used as the thank you page (CG-3).

For the hypothesis to hold there should be no overlap in the dates sessions submitted forms for CG-1, CG-2, and CG-3. In addition, the pages signup.html and thanks.html should not be visited by sessions after CG-2 and CG-3 were active, respectively. In support of the hypothesis, the final column of table 43 shows a clear separation in the dates sessions submitted forms for each of the contact goals[35]. In addition, table 44 illustrates the date ranges when each page was visited by any session only falls within the time period the contact goal was active. Therefore, it was believed that all three contact goals represented an evolution of the same contact goal, and thus they were combined into one contact goal.

[34]Indirect matches were not examined for CG-1 versus CG-3 since they do not share any common pages. Indirect matches were also not done for CG-2 versus CG-3 since the third page differed.

[35]Sessions were classified to the three contact goals according to preprocessing step 11.

Table 43: Site-centric: Conflicting Contact Goal Pages – Web site D

Contact Goal	signup.html	signup1.html	thanks.html	Classified Sessions
CG-1	1		3	9/12/07 5:50 PM – 1/28/08 7:28 PM
CG-2		1	3	1/30/08 7:44 PM – 5/15/08 9:47 PM
CG-3		1 & 3		5/19/08 3:12 PM – 9/23/08 10:52 PM

Table 44: Site-centric: Web site D Page Visitations

Page	Active Visitation Range
signup.html	9/12/07 2:47 PM – 1/30/08 7:19 PM
signup1.html	1/30/08 7:22 PM – 9/23/08 11:05 PM
thanks.html	9/12/07 6:03 PM – 5/15/08 10:23 PM[a]

[a] There were two additional visits to thanks.html after 5/15/08 on 7/12/08 and 9/16/08. However, since there were only two views during a four-month period, the page was considered inactive after 5/15/08.

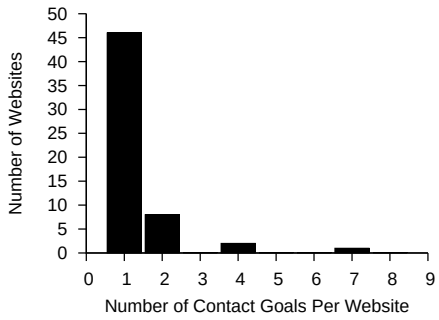


Figure 46.: Site-centric: Contact Goals Per Web site

Contact Goal Statistics A total of 77 contact goals were found on the 57 remaining Web sites in the dataset. Figure 46 illustrates how many contact goals were discovered at each Web site. The vast majority of Web sites (46) only had a single contact goal. On average, each Web site had 1.35 contact goals (0.99 standard deviation), with one site having 7 contact goals (the maximum number found on a Web site).

Step 10. Classify Goal Sessions

After discovering all the contact goals for a Web site, all goal sessions were then classified. A goal session was classified to a contact goal according to the heuristic outlined below.

Direct match – an exact pattern match without any gaps between pages. The goal session is classified to the contact goal. If no direct matches exist for any contact goal, then continue on to indirect match.

Indirect match – a pattern match of at least the first two pages, with gaps between pages allowed. The goal session is classified to the contact goal with the smallest gap (i.e., number of other pages present) between (1) the second and third page and then (2) the first and second page. This method assumed that because an automatic transfer takes place from the second to third page it is less likely another page would be visited in between that transfer. If no indirect matches exist for any contact goal, then continue on to no match.

No match – a pattern match of at least the first two pages (with or without gaps) is not found for any of the contact goals. The goal session is not classified to any contact goal[36].

[36] A session without a match is not classified to any contact goal, even on Web sites with only a single discovered

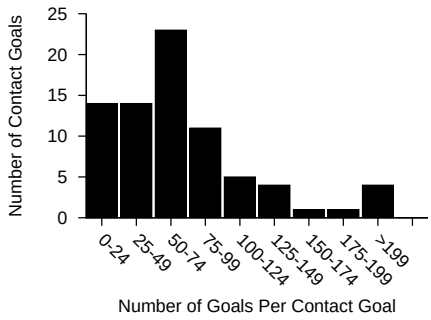


Figure 47.: Site-centric: Goals Per Contact Goal

Following the heuristic outlined above, 5,827 of the 5,975 goal sessions (97.52%) were classified to a contact goal. Of the 148 unclassified goal sessions, 124 of them (83.78%) were not classifiable because the first page of their session was the form submission page. The remaining 24 goal sessions (16.22%) may have been unclassifiable due to being a match for an undiscovered contact goal, or the user may have visited the first page of the contact goal sequence during a previous session.

Figure 47 illustrates the number of goal sessions per contact goal. Out of the 77 contact goals, 49 of them (63.64%) had 50 or more goal sessions. Table 45 displays the mean, standard deviation, minimum, and maximum number of goal sessions for all 77 contact goals.

Table 45: Site-centric: All Contact Goals Stats

	Mean	St. Dev.	Minimum	Maximum
Goals per contact goal	75.68	80.57	5	587

Step 11. Remove Web sites without any contact goals having $\geq$ 50 goal sessions

For the eleventh step, Web sites without any contact goals having at least 50 goal sessions were removed. A total of 10 Web sites were removed along with the 17 corresponding contact goals for those sites. In addition, 28,275 sessions were removed, with 525 of those being goal sessions.

Figure 48 displays the number of goal sessions per contact goal for the remaining 60 contact goals. 49 of the 60 contact goals (81.67%) had 50 or more goal sessions. Table 46 displays the

contact goal. This is because the Web site may contain other contact goals that were simply too small to detect during the previous preprocessing step.

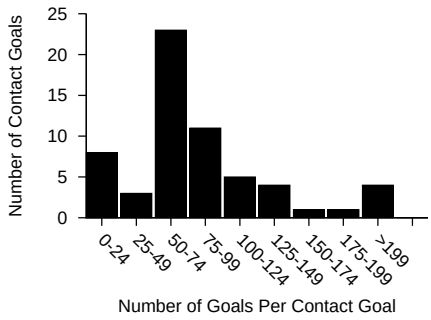


Figure 48.: Site-centric: Web sites with $\geq$ 50 Goal Sessions Per Contact Goal

mean, standard deviation, minimum, and maximum number of goal sessions for the remaining 60 contact goals.

Table 46: Site-centric: Web sites with $\geq$ 50 Goal Sessions – Contact Goals Stats

	Mean	St. Dev.	Minimum	Maximum
Goals per contact goal	88.37	86.89	5	587

Step 12. Classify other contact goal sessions as non-goal sessions

For the twelfth and final step of the process, the contact goal with the highest number of goal sessions was selected for each Web site as the contact goal to be analyzed. The selection of a single contact goal per Web site was done to simplify the analysis. Goal sessions from any other contact goal at the Web site were classified as non-goal sessions[37].

As there were 47 Web sites, a total of 47 contact goals were selected to be analyzed. The goal sessions at the remaining 13 contact goals were classified as non-goal sessions. 4,979 goals were achieved on the 47 selected contact goals (93.91%), while the remaining 323 goals from the 13 not-selected contact goals (6.09%) were classified as non-goal sessions.

[37]Even though it is known these other goal sessions did achieve a goal, the goal was for a different contact form, and thus not the goal being focused on at the Web site being analyzed.

6.2.2 Final Dataset

The following subsections provide general statistics about the final dataset, along with characteristics of the Web sites[38] and sessions in the dataset.

General Statistics

Table 47 displays general statistics for the final dataset. The first row of the table lists the total number of sessions in the dataset. Each row after the first lists the total count for the metric and also its percentage compared to the total number of sessions. The overall conversion rate of the dataset was 1.99%, which is similar to conversion rates found at e-commerce Web sites (Moe, 2003; Sismeiro and Bucklin, 2004). Unique visitors accounted for 80.69% of the sessions whereas 19.31% of the sessions were from repeat visitors. Lastly, 1,229,190 pages were viewed over all 250,162 sessions from the 47 Web sites in the dataset.

Table 47: Site-centric: Final Dataset Statistics

	n	%
Sessions	250,162	n/a
Goals sessions	4,979	1.99%
Non-goal sessions	245,183	98.01%
Unique visitors	201,845	80.69%
Repeat visits	48,317	19.31%
Pages viewed	1,229,190	n/a
Web sites	47	n/a

Web Site Characteristics

Table 48 provides the mean, standard deviation, minimum, and maximum values from all 47 Web sites for the number of days a site was active in the dataset; the number of valid and excluded Web

[38] Since there is only a single contact goal at a Web site, the terms "Web site" and "contact goal" will be used interchangeably.

pages on a site; the number of goal, non-goal, and total sessions visiting each site; and the conversion rate from each site. Valid pages included all pages flagged as valid from step number two in the preprocessing section. Excluded Web pages were those pages flagged as necessary conditions for achieving a goal. Excluded pages were removed from a session when mining patches and trails.

Table 48: Site-centric: Web Site Characteristic Statistics

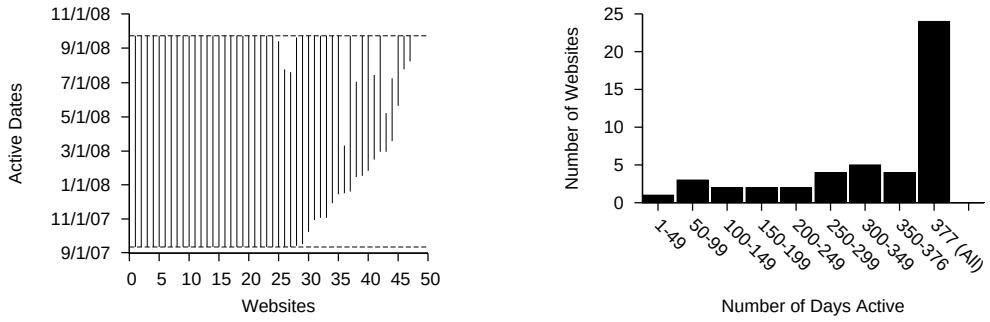
	Mean	St. Dev.	Minimum	Maximum
WEB SITE ACTIVITY				
Days Active	308.36	104.37	46	377
PAGES				
Valid pages	16.36	13.00	5	79
Excluded pages	2.04	0.29	2	4
SESSIONS				
Total sessions	5,322.60	7,473.76	245	44,405
Goal sessions	105.94	90.13	51	587
Non-goal sessions	5,216.66	7,427.53	192	44,111
OTHER				
Conversion	5.26%	5.70%	0.51%	24.25%

The entire dataset was collected over a 377 day period (09/12/2007 to 09/23/2008). On average, the 47 Web sites in the final dataset were active for 308.36 days (81.79%)[39]. One Web site was only available for roughly a month and a half (46 days), while a number of Web sites were present during the greater than one-year data collection process (377 days). The actual dates in which a Web site was active is shown in figure 49a[40], where the dashed lines indicate the beginning and ending dates of the data collection period. Figure 49b is a histogram illustrating the number of Web sites with a specified number of active days.

[39] Activity is determined by finding the first and last session visited at each Web site. There may be periods of time between the first and last session visit dates in which no activity occurred on the Web site.

[40] The Web sites were sorted in ascending order by first session date and then descending order by last session date.

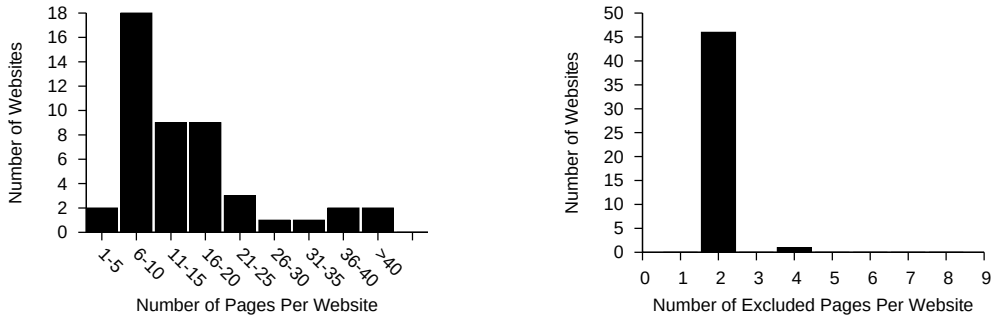
Of the 47 Web sites, 27 of them (57.45%) were active from the first day of data collection. 24 of those 27 Web sites (51.06%) remained active by the last day of data collection. For the 20 Web sites (42.55%), which were not present at the beginning of data collection, 14 of them (70.00%) were still active by the end of the data collection period.



(a) Active Duration of Web sites (b) Web sites by Active Duration

Figure 49.: Site-centric: Web sites' Activity

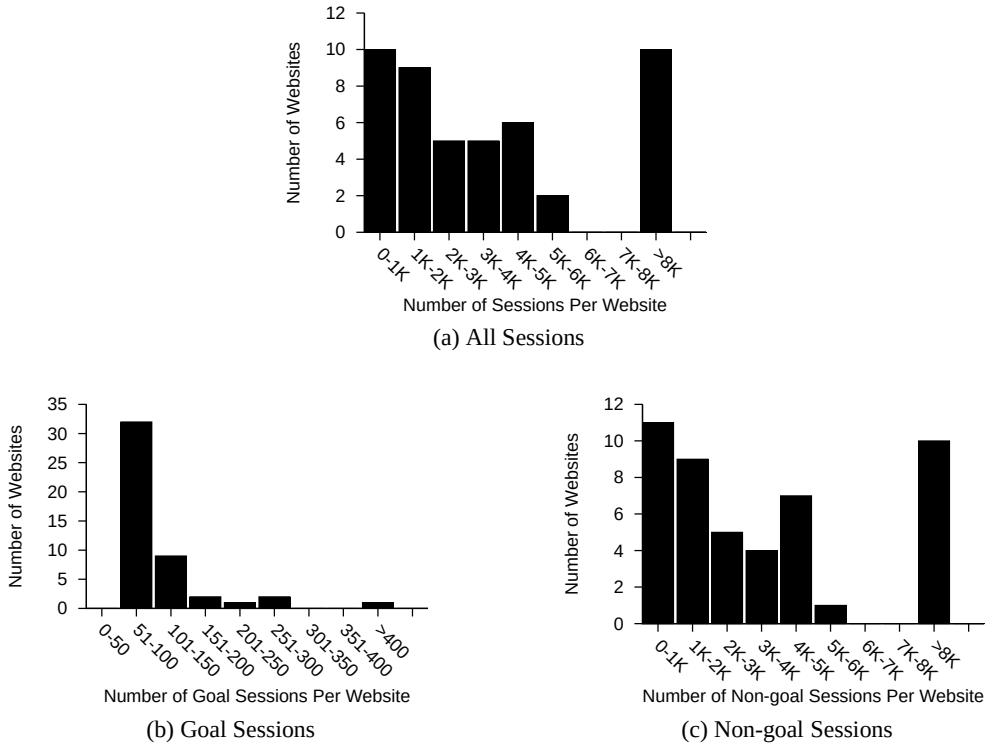
Figures 50a and 50b illustrate the distribution of number of valid and excluded Web pages for each Web site, respectively. As seen in figure 50a, most of the Web sites (31 out of 47 (70.21%)) were fairly small in size having fewer than 17 valid Web sites (16.36 Web pages on average), with the largest site having 79 pages. In terms of excluded Web pages (figure 50b), 46 of the 47 Web sites (97.87%) excluded only two Web pages. This means that the vast majority of Web sites had symmetrical contact goal patterns (e.g., contact form, form submission, contact form). The Web site which combined three contact goals together (from preprocessing step nine) was the only Web site with more than two excluded pages (the site had the maximum of four excluded pages).



(a) Valid Pages (b) Excluded Pages

Figure 50.: Site-centric: Web site Pages Histograms

On average, each Web site had 5,322.60 total sessions visiting the Web site, with almost 50 times as many non-goal sessions as goal sessions. Each Web site had, on average, 105.94 goal sessions (1.99%) and 5,216.66 non-goal sessions (98.01%). Figures 51a – 51c illustrate the distribution of the number of total, goal, and non-goal sessions for each Web site, respectively. The majority of Web sites (37 out of 47 (78.72%)) had fairly light traffic, having less than 8,000 total sessions (figure 51a). However, there were 10 Web sites (21.28%) with more than 8,000 total sessions, with the most heavily-visited site having 44,405 total sessions. In terms of goal sessions (figure 51b), 41 Web sites (87.23%) had between 50 and 150 goal sessions, with 32 of those 41 (78.05%) having 50 to 100 goal sessions.



(a) All Sessions

(b) Goal Sessions

(c) Non-goal Sessions

Figure 51.: Site-centric: Web site Sessions Histograms

The average conversion rate for the 47 Web sites was 5.26%, with one site having the highest rate of 24.25%. Figure 52 illustrates the distribution of conversion rates for each Web site. 25 of the 47 Web sites (53.19%) had less than a 3% conversion rate. 14 of the Web sites (29.79%) had between a 3% and 8% conversion rate. The remaining eight Web sites (17.02%) had a conversion rate higher than 8%.

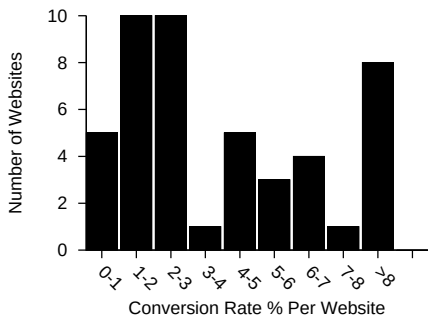


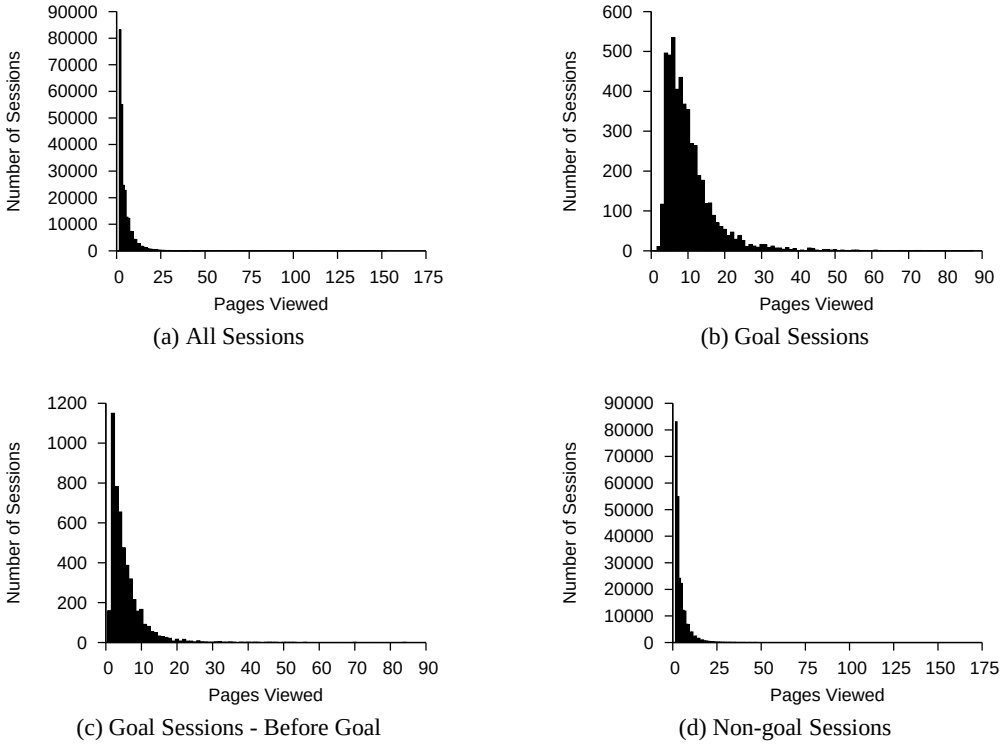
Figure 52.: Site-centric: Web site Conversion Histogram

Session Characteristics

Table 49 provides the mean, standard deviation, minimum, and maximum values for the number of pages viewed and duration from all 250,162 sessions in the dataset. For each metric, values are provided for three sets of sessions: goal, non-goal, and all sessions. Since the site-centric clickstream model of information foraging uses measures calculated prior to the submission of a contact form to predict a goal, the number of pages viewed and session duration are also provided for goal sessions at the point right *before* they submitted a contact form.

Each session consisted, on average, of less than five page views (4.91), with a maximum of 152 pages viewed by one session. Goal sessions viewed over twice as many pages per session, on average, compared to non-goal sessions (10.34 versus 4.80). Even when only the pages viewed before a form submission were considered, goal sessions still viewed almost one additional page, on average, than non-goal sessions (5.60 versus 4.80). Goal sessions also viewed a little more than half of all their pages (54.16% more pages) before submitting a contact form. Figures 53a – 53d show the distribution of pages viewed by number of sessions.

The average duration from all 250,162 sessions was 3.78 minutes, with one session spending over 134 minutes on a site. The difference between goal and non-goal session duration was even more pronounced than the number of pages viewed. Goal sessions spent over three times as many minutes on a site compared to non-goal sessions (11.46 min versus 3.62 min). Before submitting a goal, goal sessions spent over two times as much time as the average non-goal session (8.80 min versus 3.62 min). In addition, goal sessions spent more than three-quarters of their time (76.79% of their time) browsing the site before submitting a contact form. Figures 54a – 54d illustrate the distribution of session duration in minutes by number of sessions.



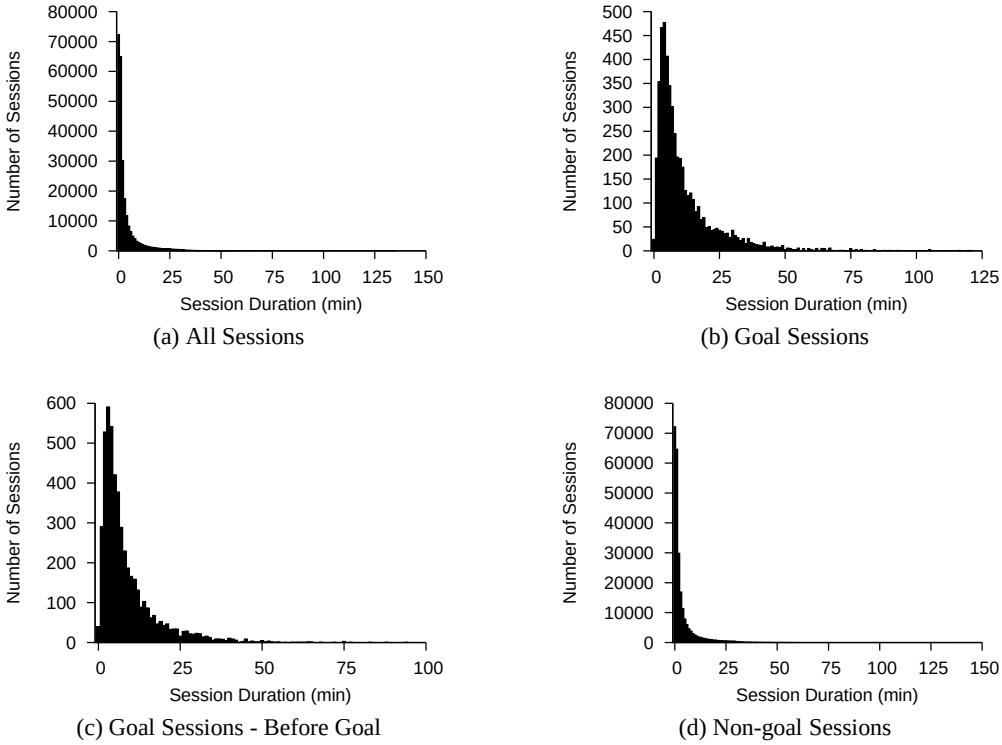
(a) All Sessions

(b) Goal Sessions

(c) Goal Sessions - Before Goal

(d) Non-goal Sessions

Figure 53.: Site-centric: Session Pages Viewed Histograms



(a) All Sessions

(b) Goal Sessions

(c) Goal Sessions - Before Goal

(d) Non-goal Sessions

Figure 54.: Site-centric: Session Duration Histograms

Table 49: Site-centric: Session Characteristic Statistics

	Mean	St. Dev.	Minimum	Maximum
PAGES VIEWED				
All sessions	4.91	5.05	2	152
Goal sessions	10.34	7.17	2	87
Before goal	5.60	5.26	1[a]	84
Non-goal sessions	4.80	4.94	2	152
SESSION DURATION (MIN)				
All sessions	3.78	6.99	0.00	134.75
Goal sessions	11.46	11.96	0.17	120.15
Before goal	8.80	9.21	0.08	94.17
Non-goal sessions	3.62	6.76	0.00	134.75

[a] A minimum of one page viewed is valid (when sessions are restricted to at least two pages) because only those pages viewed *before* the form submission were included. In this situation the contact form page was viewed first and then the form was submitted.

6.3 Conclusion

This chapter provided an overview of the datasets used to test the user- and site-centric clickstream models of information foraging. An explanation was given regarding the process by which the data was captured, along with the preprocessing steps undertaken to arrive at the final dataset for each model. General statistics were then shown for each dataset, along with the Web site and session characteristics from each dataset. Graphical representations of many metrics were also shown to illustrate the distributions of values within the datasets.

Chapter 7
Results

Presented in this chapter are the results for both the user- and site-centric clickstream models of information foraging. Descriptive statistics, checks of the assumptions for the statistical tests used to test the model's hypotheses, and the results for each hypothesis are described individually for both of the models. In addition, the site-centric section provides a sensitivity analysis of eight different mining significance and support levels used to calculate measures for the seven hypotheses that relied on learned patches and trails.

7.1 User-centric Clickstream Model of Information Foraging

The user-centric model consisted of four hypotheses regarding the value of an entire site as a patch. The descriptive statistics of the dataset, metric statistics for each hypothesis, and checks of assumptions for the three statistical tests used to test the hypotheses are presented in §7.1.1. The results from each of the statistical tests performed for all four hypotheses are then detailed in §7.1.2.

7.1.1 Descriptive Statistics

Table 50 details the mean, standard deviation, median, minimum, and maximum number of user-sessions by Web site. Statistics for goal and non-goal user-sessions at each Web site are also shown[1].

On average, each Web site had 9,545.06 user-sessions with more than 45 as many non-goal user-sessions as goal user-sessions (9,339.02 versus 206.04). The average conversion rate for each Web site (2.21%) was similar to the two percent conversion rate typically found at e-commerce sites (Moe, 2003; Sismeiro and Bucklin, 2004)[2].

Table 51 presents the mean, standard deviation, median, minimum, and maximum values of all 52 Web sites for each of the four metrics in the user-centric model. The statistics for the first two

[1] Further descriptive statistics for the dataset can be found in §6.1.2.

[2] The average conversion rate when taking the average from each Web site was 6.70% (see §6.1.2).

Table 50: User-centric: User-sessions by Site

	Mean	St. Dev.	Median	Min	Max
All	9,545.06	14,078.48	4,214.50	406	76,138
Goal	206.04	154.91	141.00	51	597
Non-goal	9,339.02	14,064.45	3,989.00	290	76,041

metrics are displayed in three groups of user-sessions: all, goal, and non-goal. The statistics for the last two metrics are also displayed for all user-sessions, but were also split to show the conversion rate within two groups of sessions: those users that returned during the same session and those foragers who stayed on the Web site during the entire session[3].

The average relative duration of users spent 1.27 more minutes at each target Web site than on other sites. The goal target sessions spent, on average, 7.44 more minutes on the target Web site than at other e-commerce sites within their respective user-sessions. The non-goal target sessions spent 4.89 *fewer* minutes on the target Web site compared to the median time spent on other Web sites. A similar distinction between goal and non-goal target sessions was also seen in the relative number of pages visited. Over 20 more pages (20.27) were visited by goal sessions at their target Web site, while non-goal sessions visited roughly one fewer page (-1.07) on their target site compared to other Web sites within a user-session.

The final two measures demonstrated the conversion rates from two groups of sessions. On average, a 0.91% increase in conversion rate was found for sessions that stayed on a Web site the entire session versus those that left and returned during the same session (7.32% versus 6.41%). A similar difference was also found between the two groups within the REPEAT measure. A 0.62% increase in conversion rate, on average, was found for sessions that were previous visitors of the Web site versus new visitors (6.97% versus 6.35%).

[3]For REPEAT, the groups demonstrated the conversion rate of sessions that had visited the Web site before and those sessions that were new visitors to the Web site.

Table 51: User-centric: Metric Statistics

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – SITE-PATCH						
RELDUR (in minutes)						
All	52	1.27	9.28	−2.00	−11.00	43.50
Goal		7.44	9.59	6.00	−10.75	43.50
Non-goal		−4.89	2.09	−5.00	−11.00	0.00
RELPGS						
All	52	9.60	15.57	2.00	−6.00	77.00
Goal		20.27	15.79	17.25	1.00	77.00
Non-goal		−1.07	2.77	−1.25	−6.00	12.00
RETURN						
All	52	6.87%	7.51%	3.88%	0.05%	41.04%
$P\,(Goal\|Return)$		6.41%	6.68%	3.78%	0.13%	24.33%
$P\,(Goal\|Stayed)$		7.32%	8.30%	4.11%	0.05%	41.04%
REPEAT						
All	52	6.66%	7.78%	3.64%	0.10%	46.63%
$P\,(Goal\|Repeat)$		6.97%	6.84%	4.33%	0.10%	28.72%
$P\,(Goal\|New)$		6.35%	8.68%	2.83%	0.20%	46.63%

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Assumptions of Statistical Tests

Table 52 lists the assumptions for each of the three statistical tests used to test the model's hypotheses[4]. A ☑ symbol indicates the assumption was met for the statistical test, while a □ symbol means the assumption was not met. If both a ☑ and a □ symbol are shown then the assumption held for some metrics, but not for all of the metrics. There were a total of five assumptions for the paired t-test (assumptions three through seven); four for the exact Wilcoxon signed rank test (assumptions three through six); and three for the dependent-samples sign test (assumptions one, two, and five).

Table 52: User-centric: Assumptions of Statistical Tests

#	Assumption	t-Test	Wilcoxon	Sign Test
1	The pairs (X_i, Y_i) are internally consistent, in that if P(+) > P(-) for one pair (X_i, Y_i), then P(+) > P(-) for all pairs.			☑
2	The measurement scale is at least ordinal within each pair.			☑
3	The measurement scale of the D_is is at least interval.	☑	☑	
4	The D_is all have the same mean.	☑	☑	
5	The D_is (or bivariate random variables (X_i, Y_i)) are mutually independent.	☑	☑	☑
6	The distribution of each D_i is symmetric.	☑□	☑□	
7	The D_is are identically distributed normal random variables.	□		

(Conover, 1999, pg. 157-158, 353, 363)

Further details about whether assumptions were met or not for each of the statistical tests are provided below. The tests are presented in order of which test had the least to most stringent assumptions: sign test, Wilcoxon test, and t-test.

Dependent-samples Sign Test

All three assumptions of the sign test were fully met.

Assumption 1 – Each observation pair was internally consistent. If P(+) > P(-), P(+) < P(-), or P(+) = P(-) for a single observation pair, then P(+) > P(-), P(+) < P(-), or P(+) = P(-) was the same across all observation pairs, respectively.

[4]Assume within the data there are n pairs of X and Y observations $(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)$. For each observation pair, the difference D_i is calculated between X_i and Y_i, where $D_i = Y_i - X_i$.

Assumption 2 – Each metric used in this research was a quantitative variable measured on at least an interval scale.

Assumption 5 – Each pair of bivariate random variables (X_i, Y_i) was taken from a different and independent Web site.

Exact Wilcoxon Signed Rank Test

Three of the four assumptions for the exact Wilcoxon signed rank test were met. The fourth assumption dealing with symmetry of the D_is was met for some of the metrics, but not for all of the metrics.

Assumption 3 – Each metric used in this research was a quantitative variable measured on at least an interval scale.

Assumption 4 – Each of the differences (D_i) was taken from a Web site within the same population. Therefore, the mean of each difference was expected to be the same.

Assumption 5 – Each of the differences (D_i) was taken from a different and independent Web site.

Assumption 6 – The last four columns from table 53 show two different measures of skewness for the four metrics: coefficient of skewness and quartile skew coefficient. Since the Wilcoxon test only considers data points with non-zero differences, only the skew values from the "No zeros" columns were analyzed[5].

[5]None of the 52 Web sites had the same median value for goal and non-goal sessions for any of the four measures. Therefore, all 52 Web sites were included when calculating skew for both the "All" and "No Zeros" columns.

Table 53: User-centric: Metric Normality and Skew

		N		Lilliefors		Shapiro		Skew		Quartile Skew	
Hyp.	Metric	Total	No Zeros	D	p-Value	W	p-Value	All	No Zeros	All	No Zeros
INFORMATION PATCH – SITE-PATCH											
SC1	RELDUR	52	52	0.22	$< 0.0001^{***}$	0.75	$< 0.0001^{***}$	−2.23	−2.23	0.00	0.00
SC2	RELPGS	52	52	0.18	0.0002^{***}	0.83	$< 0.0001^{***}$	−1.79	−1.79	0.11	0.11
SC3	RETURN	52	52	0.21	$< 0.0001^{***}$	0.69	$< 0.0001^{***}$	2.40	2.40	0.93	0.93
SC4	REPEAT	52	52	0.29	$< 0.0001^{***}$	0.65	$< 0.0001^{***}$	3.32	3.32	−0.29	−0.29

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

The first two of the skew columns provide the commonly used coefficient of skewness (g), as shown in equation 7.1 (Helsel and Hirsch, 1992). In equation 7.1, n is the number of elements, x_i is the value of the i^{th} element, $\bar{X}$ is the sample mean, and s is the sample standard deviation. Although widely used, when using the coefficient of skewness "...an otherwise symmetric distribution having one outlier will produce a large (and possibly misleading) measure of skewness" (Helsel and Hirsch, 1992, pg. 10).

$$g = \frac{n}{(n-1)*(n-2)} \sum_{i=1}^{n} \frac{\left(x_i - \bar{X}\right)^3}{s^3} \tag{7.1}$$

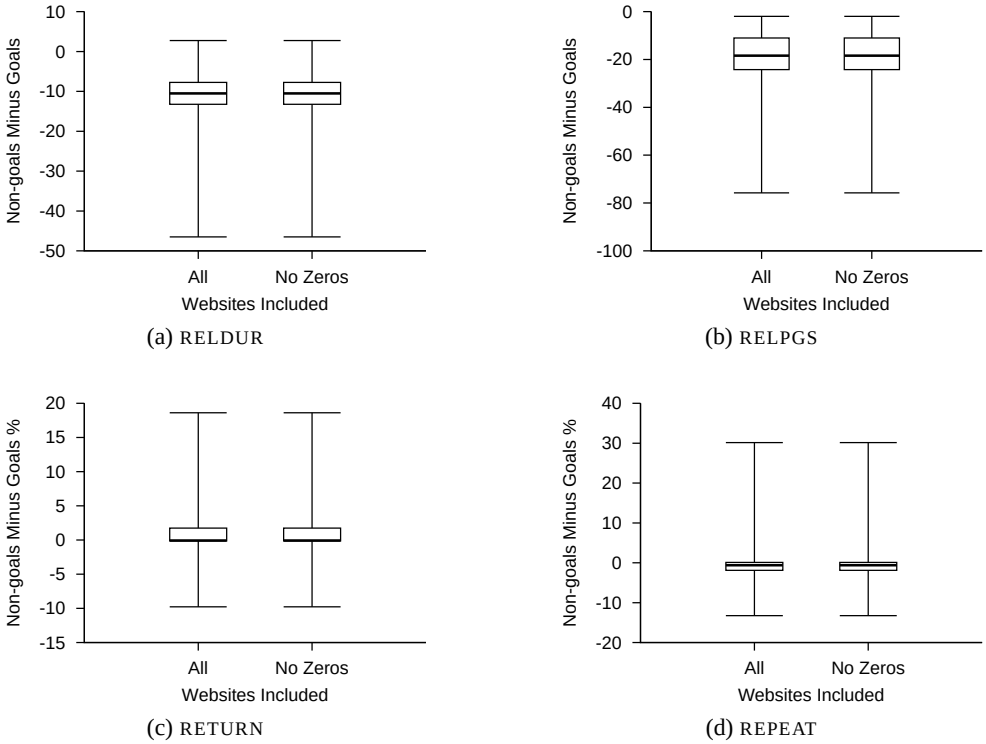
Due to the sensitivity of the coefficient of skewness to outlying points, a more robust and resilient measure of skew which is not affected by outliers was used (last two columns of table 53). The formula for the quartile skew coefficient is shown in equation 7.2 (Helsel and Hirsch, 1992), where $P_{0.25}$, $P_{0.50}$, and $P_{0.75}$ refer to the lower quartile, median, and upper quartile of the data, respectively. The quartile skew coefficient can range from negative one to one. Since the quartile skew measure only considers the difference between the upper and lower quartiles and the median, outlying points (such as the maximum and minimum) do not impact the value of the skew measure.

$$qs = \frac{(P_{0.75} - P_{0.50}) - (P_{0.50} - P_{0.25})}{P_{0.75} - P_{0.25}} \tag{7.2}$$

Besides statistics on skew as shown in table 53, figure 55 is also provided to graphically show the distribution of points for each measure.

RELDUR and RELPGS were both negatively skewed (-2.23 and -1.79), having a long tail below the median. However, between the lower and upper quartiles, the distribution of points appears to be mostly symmetric around the median. Examining the quartile skew coefficient values, RELDUR did not demonstrate any skew (0.00), while RELPGS had a slight positive skew (0.11).

RETURN and REPEAT had a skew opposite of the first two measures and were positively skewed (2.40 and 3.32). Both measures had a long tail above the median. The quartile skew coefficient demonstrated a severe positive skew for RETURN (0.93) and a moderate negative skew for RE-



(a) RELDUR (b) RELPGS

(c) RETURN (d) REPEAT

Figure 55.: User-centric: Difference Plots

PEAT (-0.29).

RELDUR was the only metric which met the assumption of no skew (using the quartile skew coefficient). RELPGS and REPEAT both had slight to moderate amounts of skew and thus did not fully meet the assumption. Lastly, RETURN was found to have a severe skew and did not meet the assumption of symmetry.

Paired t-Test

Three of the five assumptions for the paired t-test were met. The fourth assumption dealing with symmetry of the D_is was met for some of the metrics, but not for all of the metrics. The fifth assumption of normality was not formally met for any of the metrics.

Assumption 3 – Each metric used in this research was a quantitative variable measured on at least an interval scale.

Assumption 4 – Each of the differences (D_i) was taken from a Web site from the same population. Therefore, the mean of each difference was expected to be the same.

Assumption 5 – Each of the differences (D_i) was taken from a different and independent Web site.

Assumption 6 – The symmetry for each measure was determined in the same manner as for the exact Wilcoxon signed rank test[6]. Of the four measures, three of the measures had between zero and moderate amounts of skew: RELDUR did not have any skew, RELPGS had slight skew, and REPEAT had moderate skew. RETURN had severe skew and did not meet the assumption of symmetry.

Assumption 7 – Two formal statistical tests of normality were performed on the differences (D_is) for each of the metrics[7]: Lilliefors (Kolmogorov-Smirnov) and Shapiro-Wilk normality tests (Conover, 1999)[8]. Each test has a null hypothesis that the data follows a normal distribution

[6]The "All" column for values of skew was used (table 53) to determine skew for the t-test because the t-test uses all differences (non-zero and zero). Since all Web sites had a difference between goal and non-goal sessions, the "All" and "No Zeros" columns are identical.

[7]Symmetry is a necessary, but not sufficient, condition for normality. Although RETURN was severely skewed and thus not symmetrical, the tests of normality were still performed on the measure for purposes of completeness.

[8]Although presented, formal tests of normality (such as Lilliefors and Shapiro-Wilk) are known to be sensitive to even slight departures from normality (Mendenhall and Sincich, 2003).

with an unspecified mean and variance. Lilliefors is a non-parametric normality test, whereas Shapiro-Wilk has been found to have greater power than other tests (such as Lilliefors) in many situations (Conover, 1999).

All four measures rejected the null hypothesis of a normal distribution using both the Lilliefors and Shapiro-Wilks tests. In addition to the tests of normality, the skew values from table 53 and the graphical depiction of each metric's points (figure 55) provided further evidence that the measures did not follow a normal distribution.

Overall, only the assumptions for the dependent-samples sign test were fully met. Therefore, the sign test was used to test the hypotheses of the user-centric model[9]. The assumptions for the Wilcoxon test and t-test were not completely met and are provided only for comparison purposes. The lack of symmetry (and normality) for some of the measures means the results from the Wilcoxon and t-test must be interpreted with caution.

7.1.2 Hypotheses Testing

Table 54 presents the results for the four user-centric hypotheses. The first two columns of the table list the hypothesis number and name of the metric being tested. The third and fourth columns list the total number of Web sites and the number of sites with a non-zero difference (i.e., $D_i \neq 0$), respectively. The total number of Web sites was used in the t-test, while only Web sites with non-zero differences were used for the Wilcoxon and sign tests[10]. Columns five through seven list the t statistic, degrees of freedom (df), and p-value for the t-test. The eighth and ninth columns display the V statistic and p-value for the Wilcoxon test. The final two columns list the S statistic and p-value for the sign test[11].

[9] The sign test is generally the least powerful of the three tests (Conover, 1999). However, as all of the sign test's assumptions were met, greater confidence can be given to the results of the sign test compared to the other two tests.

[10] Within the user-centric dataset all of the Web sites had non-zero differences.

[11] Results of the sign test are presented below since all three assumptions for the test were met. The results of the t-test and Wilcoxon test are provided in footnotes. Since neither the t-test nor the Wilcoxon test met all of their assumptions, the results of those tests should be interpreted with caution.

Table 54: User-centric: Results

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – SITE-PATCH										
UC1	RELDUR	52	52	9.71	51	$< 0.0001^{***}$	1,376	$< 0.0001^{***}$	51	$< 0.0001^{***}$
UC2	RELPGS	52	52	10.37	51	$< 0.0001^{***}$	1,378	$< 0.0001^{***}$	52	$< 0.0001^{***}$
UC3	RETURN	52	52	−1.96	51	0.9724	445	0.9874	22	0.8942
UC3	(opp)[a]	52	52	1.96	51	0.0276^{**}	933	0.0129^{**}	30	0.1659
UC4	REPEAT	52	52	0.82	51	0.2075	1,021	0.0010^{***}	36	0.0039^{***}

[a] Hypothesis tested in opposite direction as original – i.e., leaving and returning will be *negatively* associated with achieving a goal on this long tail Web site.
*p ≤ 0.10; **p ≤ 0.05; ***p ≤ 0.01

UC1 – RELDUR

The first hypothesis conjectured that goal achieving foragers would spend relatively more time on the Web site where a purchase was made than on any other site they visited. The rationale behind the hypothesis was because foragers are assumed to be rational and thus are looking to reduce their search costs (Pirolli, 2007), they will only spend time on a site as long as they are obtaining value from that site (i.e., satisficing on rate of information gain (Pirolli, 2007; Simon, 1956)). Thus, more information can be assumed to be gathered on a Web site where more time is spent than other Web sites, which brings a forager one step closer to being at a point to make a decision to purchase or not.

The results of the sign test provide support for the hypothesis at α = 0.01 (S = 51; p-value = < 0.0001)[12]. Of the 52 Web sites in the dataset, 51 of them had a higher median relative duration amongst goal sessions than non-goal sessions. Compared to other sites in their user-sessions, goal sessions spent over seven additional minutes on the target Web site while non-goal sessions spent almost five minutes less.

The use of duration to explain choice behavior has not found consistent results in prior literature. The role of absolute duration in predicting choice behavior has found mixed associations (Sismeiro and Bucklin, 2004) and also differences in significance (Padmanabhan et al., 2001) depending on the task examined and data used. The results of hypothesis UC1 provide additional support for a positive association between duration and goal achievement. However, the hypothesis only supports the notion of a positive association for *relative* rather than absolute duration.

UC2 – RELPGS

The second hypothesis was similar to the first hypothesis because it also relied on the concept of satisficing (Pirolli, 2007; Simon, 1956). However, the number of pages viewed by a forager was examined instead of the duration spent at the site. Whenever a forager clicked on a link at a site that was an implicit signal the user believed other information of value would be obtained from the site. Therefore, more pages (relative to other sites) should be an indication of a greater wealth of information being obtained.

[12]Hypothesis UC1 was also significant at α = 0.01 for both the t-test (t = 9.71; df = 51; p-value = < 0.0001) and Wilcoxon test (V = 1,376; p-value = < 0.0001).

The results from the sign test supported the hypothesis at α = 0.01 (S = 52; p-value = < 0.0001)[13]. All 52 Web sites in the dataset had a higher median relative number of pages viewed for goal sessions versus non-goal sessions. Goal sessions viewed, on average, over twenty additional pages on the target Web site compared to other sites, whereas non-goal sessions viewed one fewer page than at other Web sites.

Support has been mixed in prior literature for the role number of pages viewed has on choice behavior. Absolute number of pages viewed has found positive association (Awad et al., 2006; Moe, 2003), no association (Chatterjee et al., 2003), and mixed association depending on the task (Sismeiro and Bucklin, 2004) or type of pages viewed (Van den Poel and Buckinx, 2005). Like duration, the result of this hypothesis also lends additional support to the notion of a positive association between number of pages viewed and goal achievement. However, the support is restricted to a *relative* examination of pages viewed rather than the absolute value commonly used in prior research.

Although both UC1 and UC2 were supported at α = 0.01, RELPGS was slightly better at distinguishing between the two groups of sessions (S = 52 versus 51). However, part or all of the difference between the two hypotheses may have been an artifact of measurement constraints, since RELDUR was only measured at the minute-level. Thus, a finer-grained measurement may be better able to tease out differences in duration between sessions than what was shown in the user-centric model[14].

UC3 – RETURN

The third hypothesis examined the returning behavior of a forager. In particular, it was hypothesized that foragers who returned during the same session would be more likely to achieve a goal than foragers that did not. The rationale was that users initially left a site because they expected to find another Web site with a higher rate of information gain (i.e., they followed the patch-leaving rule from the marginal value theorem (Pirolli, 2007; Charnov, 1976)). However, after the forager

[13] Hypothesis UC2 was also significant at α = 0.01 for both the t-test (t = 10.37; df = 51; p-value = < 0.0001) and Wilcoxon test (V = 1,378; p-value = < 0.0001).

[14] The site-centric model does indicate duration is a better manner of distinguishing between types of sessions than pages viewed (see §7.2.2). However, considering a different dataset was used, the results are not directly comparable. For example, the Web sites in the site-centric dataset may have had fewer pages and thus number of pages viewed would be less able to distinguish between groups of sessions. In addition, the site-centric model does not take into account behavior relative to other Web sites.

explored other aspects of their environment, they better recognized the value of the site they initially left. Therefore, a forager that returned to the site they left not only had knowledge of what was on the site, but also an expectation that the Web site would result in additional information gain, which was hypothesized to indicate greater likelihood of goal achieving behavior.

The sign test failed to support the hypothesis at any of the tested α levels (S = 22; p-value = 0.8942)[15]. Only 22 of the 52 Web sites found a greater incidence of goal achievement among sessions that left and returned as opposed to those sessions that stayed on the site for the entire session. Not only was UC3 not supported, but the expected association of returning behavior to goal achievement appeared to be incorrect. Instead of being a positive association, the results pointed toward a strong (but non-significant) negative association, i.e., a forager was *less* likely to achieve a goal if the user left a Web site and then returned within the same session (the opposite of hypothesis UC3)[16].

Although the results of UC3 were not expected, they did provide additional support highlighting the efficacy of a forager's ability to search with only imperfect information and limited computational resources. For example, foragers appeared to be capable of judging the rate of information gain and value of a Web site relatively well, according to their need. The efficacy of foragers' search behavior was informally backed up because more users who purchased a product from their target Web site did not feel the need to visit other Web sites during their session[17].

As far as can be determined, prior literature has not examined the returning behavior of a user during the same session. Therefore, the results of this hypothesis provide an initial (but non-significant) clue into the relationship between returning behavior during the same session and goal achievement.

[15]Hypothesis UC3 was also not supported at any of the tested α levels for both the t-test (t = -1.96; df = 51; p-value = 0.9724) and Wilcoxon test (V = 445; p-value = 0.9874).

[16]The opposite of hypothesis UC3 was also not supported at any of the tested α levels (S = 30; p-value = 0.1659). However, the opposite of hypothesis UC3 *was* supported at α = 0.05 for both the t-test (t = 1.96; df = 51; p-value = 0.0276) and Wilcoxon test (V = 933; p-value = 0.0129). The discrepancy of findings may be a symptom of the sign test's lack of power (the rank or actual differences between data points are not used in the sign test). However, another possibility may be the t-test and Wilcoxon test are providing inaccurate results, especially when considering the extreme skew of the RETURN measure (quartile skew of 0.93).

[17]This assumes the action of purchasing a product from the target Web site was a "good" decision.

UC4 – REPEAT

The final hypothesis also examined returning behavior, but did so by looking at how past visitations of a Web site affected the propensity of foragers to achieve a goal. The expectation was prior visitation of valuable sites would stand out more in a person's memory (i.e., be more easily accessible) than less valuable sites. Thus, a repeat visitor would be more likely to achieve a goal because of the expectation that the user was familiar with the site and had an understanding of the available information from the site.

Using the sign test, the final hypothesis was supported at α = 0.01 (S = 36; p-value = 0.0039)[18]. 36 of the 52 Web sites had a higher median probability of a purchase amongst goal sessions when a user had visited the site before.

Prior literature has found mixed associations (dependent on the task) between users returning during different sessions and completing a task (Sismeiro and Bucklin, 2004). The results of hypothesis UC4 lends additional support of a positive association between repeat visitation behavior and goal achievement.

Summary of Results

Table 55 summarizes the results of the hypotheses testing. Of the four hypotheses, UC1, UC2, and UC4 were all supported at α = 0.01. Hypothesis UC3 was not supported in either its original or opposite form at any of the tested alpha levels (0.01, 0.05, or 0.10).

[18]Hypothesis UC4 was not significant at α = 0.10 for the t-test (t = 0.82; df = 51; p-value = 0.2075), but was significant at α = 0.01 for the Wilcoxon test (V = 1,021; p-value = 0.0010). The t-test may have failed to reach significance because the actual difference between the goal and non-goal sessions was only 0.62% (6.97% versus 6.35%). The Wilcoxon and sign tests do not consider the absolute difference, but rather the relative difference (i.e., rank) or if one group was higher than the other.

Table 55: User-centric: Hypotheses Results Summary

Hyp.	Metric	Hypothesis Supported?
INFORMATION PATCH – SITE-PATCH		
UC1	RELDUR	Yes***
UC2	RELPGS	Yes***
UC3	RETURN	No
UC3	(opp)[a]	No
UC4	REPEAT	Yes***

[a] Hypothesis tested in opposite direction as original – i.e., leaving and returning will be *negatively* associated with achieving a goal on this long tail Web site.
*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

7.2 Site-centric Clickstream Model of Information Foraging

The site-centric model consisted of nine hypotheses that were concerned with both information scent and patches. Descriptive statistics of the dataset and each measure along with checks of assumptions for the three statistical tests used to test the hypotheses are presented in §7.2.1. The results of all nine hypotheses are then detailed in §7.2.2. As three of the hypotheses (seven total measures) used metrics derived from learning patches and trails, a sensitivity analysis was performed at eight different mining levels of significance and support. The descriptive statistics and results of the sensitivity analysis are provided in §7.2.3.

7.2.1 Descriptive Statistics

Table 56 details the mean, standard deviation, median, minimum, and maximum number of sessions by Web site. Statistics for goal and non-goal sessions at each Web site are also shown[19]. The data is separated in table 56 into three groups: the entire dataset, training set, and testing set.

The entire dataset was used to test the six hypotheses which did not rely on mining patches or trails (SC1–SC4, and SC7–SC8). The training dataset was used to discover patches and trails that would eventually be used to calculate measures for hypotheses SC5a-c, SC6, and SC9a-c[20]. The

[19] Further descriptive statistics for the dataset can be found in §6.2.2.

[20] The training set consisted of the first 70% of goal sessions (and all non-goal sessions occurring before the last goal

Table 56: Site-centric: Sessions by Site

	Mean	St. Dev.	Median	Min	Max
ENTIRE DATASET					
All	5,322.60	7,473.76	2,637.00	245	44,405
Goal	105.94	90.13	79.00	51	587
Non-goal	5,216.66	7,427.53	2,566.00	192	44,111
TRAINING SET					
All	3,744.23	5,418.42	1,696.00	168	31,730
Goal	74.28	63.07	56.00	36	411
Non-goal	3,669.96	5,386.00	1,656.00	130	31,525
TESTING SET					
All	1,578.36	2,156.26	901.00	48	12,675
Goal	31.66	27.07	23.00	15	176
Non-goal	1,546.70	2,143.14	884.00	29	12,586

actual calculation of the measures for hypotheses SC5a-c, SC6, and SC9a-c were done using sessions from the testing set of data.

On average, each Web site had a total of 5,322.60 sessions with more than 49 as many non-goal sessions as goal sessions (5,216.66 versus 105.94). The average conversion rate for each Web site (1.99%) was similar to the two percent conversion rate typically found at e-commerce Web sites (Moe, 2003; Sismeiro and Bucklin, 2004)[21].

Overall, the training set of data represented 70.35% of all sessions. For mining purposes, each Web site had an average of 3,744.23 sessions. The training set had a similar ratio of goal versus non-goal sessions (49 times more non-goal than goal sessions) and a slightly lower conversion rate than what was seen in the entire dataset (1.98% versus 1.99%).

The testing set was also very similar in makeup to the entire dataset. Each Web site had an average of 1,578.36 sessions, with almost 49 as many non-goal sessions as goal sessions (1,546.70 versus 31.66). The conversion rate was also very similar to the entire dataset (2.01% versus 1.99%).

As seen in table 56, the makeup of the training and testing sets do not appear to differ drasti-

session added to the training set).

[21]The average conversion rate when taking the average from each Web site was 5.26% (see §6.2.2).

cally from the entire dataset. Therefore, the results of mining patches and trails and the calculation of measures using the testing and training datasets are assumed to be generalizable to the entire dataset (i.e., the results are not an artifact of the manner in which the data was split).

Table 57 presents the mean, standard deviation, median, minimum, and maximum values from all 47 Web sites for each of the six metrics that did not rely on mining patches and trails. The statistics for the first and last two metrics are displayed in three groups of sessions: all, goal, and non-goal. The statistics for the middle two metrics are also displayed for all sessions, but were split to show the conversion rate within two groups of sessions: those foragers that returned during the same session and those users who stayed on the Web site during the entire session[22].

The average duration of users at each Web site was 3.69 minutes. The goal sessions spent, on average, 4.60 more minutes on a Web site compared to the non-goal sessions (5.99 minutes versus 1.39 minutes). A similar, but less large, of a difference was also seen between the number of pages viewed between goal and non-goal sessions. On average, goal sessions viewed 0.43 more page than non-goal sessions did (4.28 versus 3.85)[23].

The middle two measures (RETURN and REPEAT) demonstrate the conversion rates from two groups of sessions. On average, a 5.53% increase in conversion rate was found for sessions that stayed on a Web site the entire session versus those that left and returned during the same session (7.24% versus 1.71%). A similar, but not as severe, difference was also found between the two groups within the REPEAT measure. A 1.02% increase in conversion rate, on average, was found for sessions that were previous visitors of the Web site versus new visitors (6.09% versus 5.07%).

The percentage of unique pages viewed, on average, was 79.42%. Goal sessions had a 20.22% increase in unique pages viewed over non-goal sessions (89.53% versus 69.31%). The difference in clickstream linearity was also similar to the percentage of unique pages viewed in both direction and difference. A 0.23 increase in clickstream linearity was seen between the goal and non-goal sessions (0.86 versus 0.75).

Table 58 lists the mean, standard deviation, median, minimum, and maximum values for the seven metrics (hypotheses SC5a-c, SC6, and SC9a-c) that were calculated from mined patches and trails at the 0.05 significance level[24]. The statistics for the metrics are displayed in three groups of

[22]For REPEAT, the groups demonstrated the conversion rate of sessions that had visited the Web site before and those sessions that were new visitors to the site.

[23]The duration and number of pages viewed for goal sessions only includes activity *before* any form submission.

[24]The use of α = 0.05 for learning patches and trails was motivated by prior research on contrast sets (Bay and Paz-

Table 57: Site-centric: Metric Statistics

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – SITE-PATCH						
SITEDUR (in minutes)						
All	47	3.69	2.87	2.89	0.33	13.32
Goal		5.99	2.32	5.90	1.39	13.32
Non-goal		1.39	0.66	1.23	0.33	3.15
SITEPGS						
All	47	4.06	1.26	4.00	2	9
Goal		4.28	1.46	4.00	2	9
Non-goal		3.85	1.00	4.00	2	7
RETURN						
All	47	4.47%	6.04%	2.16%	0.00%	31.72%
$P\,(Goal\|Return)$		1.71%	2.68%	0.81%	0.00%	14.02%
$P\,(Goal\|Stayed)$		7.24%	7.14%	4.71%	0.65%	31.72%
REPEAT						
All	47	5.58%	5.99%	3.20%	0.35%	27.27%
$P\,(Goal\|Repeat)$		6.09%	6.33%	3.55%	0.35%	27.27%
$P\,(Goal\|New)$		5.07%	5.66%	2.96%	0.46%	24.89%
STRICT INFORMATION SCENT						
UNIQUE						
All	47	79.42%	15.23%	77.50%	50.00%	100.00%
Goal		89.53%	11.94%	100.00%	58.33%	100.00%
Non-goal		69.31%	10.85%	66.67%	50.00%	100.00%
LINEAR						
All	47	0.86	0.29	1.00	0.00	1.00
Goal		0.98	0.08	1.00	0.60	1.00
Non-goal		0.75	0.37	1.00	0.00	1.00

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

sessions: all, goal, and non-goal. The statistics for the four metrics regarding page-patches were calculated from the 14 Web sites that discovered patches, while the three trail measures were calculated from the 10 Web sites that discovered trails.

The first three patch measures (PATCHMAX, PATCHLAST, and PATCHSUM) had average patch values of 0.32, 0.29, and 0.79 among all sessions, respectively. The difference between goal and non-goal sessions was similar for PATCHMAX and PATCHLAST. PATCHMAX had a difference of 0.44 (0.54 versus 0.10) and the difference for PATCHLAST was 0.37 (0.47 versus 0.10). PATCHSUM had the largest difference of the three measures with a value of 1.23 (1.40 versus 0.17), almost three times as great of a difference as either PATCHMAX or PATCHLAST.

The average duration users spent within patches was a little over one minute (68.28 seconds). Considering the average user spent 3.69 minutes on an entire site, foragers spent almost a third of their time (30.84%) within patches. Goal sessions foraged within patches, on average, 42.41 more seconds compared to non-goal sessions (89.48 seconds versus 47.07 seconds).

Unlike the patch visitation measures, the trail following measures all had similar means: 0.27, 0.26, and 0.33. Likewise, the difference between goal and non-goal sessions was also relatively similar between the three measures. The difference for TRAILMAX was 0.45 (0.50 versus 0.05), TRAILLAST was 0.43 (0.48 versus 0.05), and TRAILSUM had the largest difference of 0.56 (0.61 versus 0.05).

Patch and Trail Descriptive Statistics

Table 59 provides the mean, standard deviation, median, minimum, and maximum values for a number of descriptive measures about the learned patches and trails: number, size, coverage and value. In addition, statistics are also provided about how many patches were visited and trails were followed by foragers.

An average of 11.93 patches was discovered on 14 Web sites using the 0.05 significance mining level. Although almost 12 patches were discovered on average, there was a fairly large spread of discovered patches, with one Web site only finding a single patch and another site discovering 111 patches. In general, discovered patches were fairly small in size, consisting of only 1.82 pages. The small patch size indicates a number of valuable patches were simply individual pages on a

zani, 1999).

Table 58: Site-centric: Metric Statistics (Significant – 0.05)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	14	0.32	0.40	0.00	0.00	1.31
Goal		0.54	0.43	0.60	0.00	1.31
Non-goal		0.10	0.21	0.00	0.00	0.57
PATCHLAST						
All	14	0.29	0.34	0.00	0.00	1.03
Goal		0.47	0.36	0.47	0.00	1.03
Non-goal		0.10	0.20	0.00	0.00	0.51
PATCHSUM						
All	14	0.79	1.25	0.00	0.00	4.53
Goal		1.40	1.52	1.06	0.00	4.53
Non-goal		0.17	0.36	0.00	0.00	1.04
PATCHDUR (in seconds)						
All	14	68.28	47.02	51.38	19.00	178.00
Goal		89.48	53.06	76.63	19.00	178.00
Non-goal		47.07	28.43	38.88	20.00	134.00
RELAXED INFORMATION SCENT						
TRAILMAX						
All	10	0.27	0.37	0.00	0.00	1.04
Goal		0.50	0.40	0.52	0.00	1.04
Non-goal		0.05	0.16	0.00	0.00	0.50
TRAILLAST						
All	10	0.26	0.35	0.00	0.00	1.04
Goal		0.48	0.37	0.52	0.00	1.04
Non-goal		0.05	0.16	0.00	0.00	0.50
TRAILSUM						
All	10	0.33	0.50	0.00	0.00	1.80
Goal		0.61	0.58	0.52	0.00	1.80
Non-goal		0.05	0.16	0.00	0.00	0.50

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 59: Site-centric: Patch and Trail Metric Statistics (Significant – 0.05)

	N	Mean	St. Dev.	Median	Min	Max
PATCHES						
Number of patches	14	11.93	28.74	3.50	1	111
Patch size	14	1.82	0.64	2.00	1.00	3.00
Patch coverage	14	28.63%	13.85%	26.79%	10.00%	50.00%
Patch value	14	0.67	0.16	0.67	0.28	0.88
Patch visitation						
All	14	2.75	2.25	2.00	1.00	11.00
Goal		3.50	2.93	2.00	1.00	11.00
Non-goal		2.00	0.88	2.00	1.00	4.00
TRAILS						
Number of trails	10	4.70	8.04	1.50	1	27
Trail size	10	2.15	0.34	2.00	2.00	3.00
Trail coverage	10	26.47%	14.80%	27.44%	6.90%	50.00%
Trail value	10	0.79	0.19	0.82	0.46	1.05
Trail following						
All	10	1.60	1.14	1.00	1.00	5.00
Goal		1.80	1.48	1.00	1.00	5.00
Non-goal		1.40	0.70	1.00	1.00	3.00

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Web site. However, even though patches were small in size, they represented over a quarter of all Web pages on a Web site (28.63%).

The average value of a patch was relatively high at 0.67, indicating patches were reasonably capable of separating areas predominately visited by goal sessions versus non-goal sessions. However, the value of patches had a moderately large 0.60 point spread between the minimum (0.28) and maximum (0.88) valued patches.

Foragers visited, on average, 2.75 patches during a session, which represented 23.05% of all available patches on a site. Goal sessions visited 1.50 more patches than non-goal sessions did during a session (3.50 versus 2.00). Although non-goal sessions appeared to visit a number of patches, the statistics in table 59 only include sessions that visited *at least* one patch. Therefore, when considering all sessions, less than half of all non-goal sessions visited patches while over half of the goal sessions did visit patches (see median values for PATCHMAX, PATCHLAST, or PATCHSUM in table 58).

Valuable trails were more difficult to discover than patches, as evidenced by both the lower number of Web sites with trails (10 versus 14) and the mean number of trails discovered at each Web site (4.70 versus 11.93). Discovered trails consisted of either two- or three-page sequences (average of 2.15), which represented 26.47% of all Web pages on a Web site[25].

Although more difficult to find, trails were 0.12 points more valuable on average than patches (0.79 versus 0.67). The value of trails had a 0.59 point spread between the minimum (0.46) and the maximum (1.05) valued trails, which was only one tenth of a point lower than patches.

In terms of usage of valuable trails, foragers followed an average of 1.60 trails during a session, which represented 34.04% of all available trails on a Web site. Like patches, goal sessions also followed more trails than non-goal sessions (1.80 versus 1.40). Although fewer trails were followed in absolute terms compared to the number of patches visited, percentage-wise goal sessions followed a greater proportion of available trails on a Web site than patches (38.30% versus 29.34%).

Examples of Patches and Trails

[25]Trails may contain the same page being visited multiple times unlike patches which only represent unique pages. Thus, the percentage of trail coverage (26.47%) can still be lower than patch coverage (28.63%) even when the mean trail size (2.15) is higher than the average patch size (1.82). The preceding explanation assumed the difference in coverage was not due to dissimilar Web sites with different number of Web pages being included in the calculation.

Table 60 and 61 each present three examples of discovered patches and trails, respectively. Each table lists an identifier for the Web site along with a short description of what the purpose of the Web site was. The value of the example patch or trail along with the pages that make up the patch or trail is also provided. For patches, the order in which the pages are displayed in table 60 does *not* matter. For trails, the order of pages in table 61 *does* matter.

Table 60: Site-centric: Example Patches

Web site			
Id	Description	Support	Patch Pages[a]
1	Sell and service light-weight outboard motors	0.31	Index Outboard motor products Accessories Warrenties
2	Hair and make-up services for weddings	0.70	Hair prices Hair style examples Services offered Makeup prices
3	Small dog breeder	0.61	Index Photo album of puppies Available puppies

Examples are from a variety of different significance / support levels

[a] Order of pages does *not* matter

The first Web site from table 60 demonstrates a four-page patch of relatively low value. The patch may be of interest to a forager who had a question about the warranty coverage of outboard motors and their accessories. The second example represented a higher-valued patch than the first example. The four-page patch dealt with wedding hair and make-up services and may have been visited by an individual interested in booking the Web site owner for their wedding. Finally, the third example illustrates a patch that a forager may visit if they were interested in adopting puppies from a small dog breeder.

The first trail shown from table 61 demonstrates a moderately-valued three-page sequence. The example illustrates a trail followed when foragers are interested in learning how to deal cards. First

Table 61: Site-centric: Example Trails

Web site			
Id	Description	Support	Trail Pages[a]
4	Teaches professional card dealing	0.41	Index
			Testimonials
			Calendar of classes
5	Cosmetology school	0.54	General information
			Financial assistance
			Courses offered
6	Small financial company	0.31	Index
			Getting loans with poor credit
			Index
			Testimonials

Examples are from a variety of different significance / support levels
[a] Order of pages *does* matter

the user visited the index page, then read the posted testimonials, and finally viewed when classes were held. The second example demonstrates a likely path a potential student may follow when interested in enrolling in cosmetology school. General information about the school was read first, followed by information about available financial assistance, and finally what courses were offered at the school. The final example shows a trail where a forager retraced their steps. The index page was visited first and then again after the forager read information on how to obtain a loan with poor credit. The reason for the backtracking is not known, although it may be the navigation on the site followed a hub and spoke topology that required backtracking (i.e., all pages linked from the index page, but not to one another).

Assumptions of Statistical Tests

Table 62 lists the assumptions for each of the three statistical tests used to test the site-centric hypotheses[26]. A ☑ symbol indicates the assumption was met for the statistical test, while a □ sym-

[26] Assume within the data there are n pairs of X and Y observations $(X_0, Y_0), (X_1, Y_1), \ldots, (X_n, Y_n)$. For each observation pair, the difference D_i is calculated between X_i and Y_i, where $D_i = Y_i - X_i$.

bol means the assumption was not met. If both a ☑ and a □ symbol are shown then the assumption held for some metrics, but not for all of the metrics. There were a total of five assumptions for the paired t-test (assumptions three through seven); four for the exact Wilcoxon signed rank test (assumptions three through six); and three for the dependent-samples sign test (assumptions one, two, and five).

Table 62: Site-centric: Assumptions of Statistical Tests

#	Assumption	t-Test	Wilcoxon	Sign Test
1	The pairs (X_i, Y_i) are internally consistent, in that if P(+) > P(-) for one pair (X_i, Y_i), then P(+) > P(-) for all pairs.			☑
2	The measurement scale is at least ordinal within each pair.			☑
3	The measurement scale of the D_is is at least interval.	☑	☑	
4	The D_is all have the same mean.	☑	☑	
5	The D_is (or bivariate random variables (X_i, Y_i)) are mutually independent.	☑	☑	☑
6	The distribution of each D_i is symmetric.	☑□	☑□	
7	The D_is are identically distributed normal random variables.	☑□		

(Conover, 1999, pg. 157-158, 353, 363)

Further details about whether assumptions were met or not for each of the statistical tests are provided below. The tests are presented in order of which test had the least to most stringent assumptions: sign test, Wilcoxon test, and t-test.

Dependent-samples Sign Test

All three assumptions of the sign test were fully met.

Assumption 1 – Each observation pair was internally consistent. If P(+) > P(-), P(+) < P(-), or P(+) = P(-) for a single observation pair, then P(+) > P(-), P(+) < P(-), or P(+) = P(-) was the same across all observation pairs, respectively.

Assumption 2 – Each metric used in this research was a quantitative variable measured on at least an interval scale.

Assumption 5 – Each pair of bivariate random variables (X_i, Y_i) was taken from a different and independent Web site.

Exact Wilcoxon Signed Rank Test

Three of the four assumptions for the exact Wilcoxon signed rank test were met. The fourth assumption dealing with symmetry of the D_is was met for some of the metrics, but not for all of the metrics.

Assumption 3 – Each metric used in this research was a quantitative variable measured on at least an interval scale.

Assumption 4 – Each of the differences (D_i) was taken from a Web site from the same population. Therefore, the mean of each difference was expected to be the same.

Assumption 5 – Each of the differences (D_i) was taken from a different and independent Web site.

Assumption 6 – The last two columns from tables 63 and 64 show the quartile skew coefficient[27] for all 13 metrics[28]. Since the Wilcoxon test only considers non-zero differences, only the skew values from the "No zeros" columns were analyzed.

[27] A description of quartile skew coefficient and why it was analyzed over the traditionally used coefficient of skewness can be found in §7.1.1. The values for the coefficient of skewness are provided in tables 63 and 64 for reference purposes.

[28] The six measures that did not require mining for their calculation are shown in table 63. The remaining seven metrics that were calculated from mined patches and trails are displayed in table 64.

Table 63: Site-centric: Metric Normality and Skew

		N		Lilliefors		Shapiro		Skew		Quartile Skew	
Hyp.	Metric	Total	No Zeros	D	p-Value	W	p-Value	All	No Zeros	All	No Zeros
INFORMATION PATCH – SITE-PATCH											
SC1	SITEDUR	47	47	0.14	0.0184^{**}	0.92	0.0035^{***}	−1.15	−1.15	0.15	0.15
SC2	SITEPGS	47	28	0.23	$< 0.0001^{***}$	0.93	0.0076^{***}	−0.34	0.27	−1.00	0.33
SC3	RETURN	47	47	0.22	$< 0.0001^{***}$	0.75	$< 0.0001^{***}$	2.10	2.10	0.18	0.18
SC4	REPEAT	47	47	0.16	0.0035^{***}	0.87	$< 0.0001^{***}$	−1.30	−1.30	−0.18	−0.18
STRICT INFORMATION SCENT											
SC7	UNIQUE	47	44	0.09	0.3986	0.96	0.1368	−0.34	−0.37	−0.43	−0.26
SC8	LINEAR	47	18	0.36	$< 0.0001^{***}$	0.67	$< 0.0001^{***}$	−1.26	0.21	−1.00	0.07

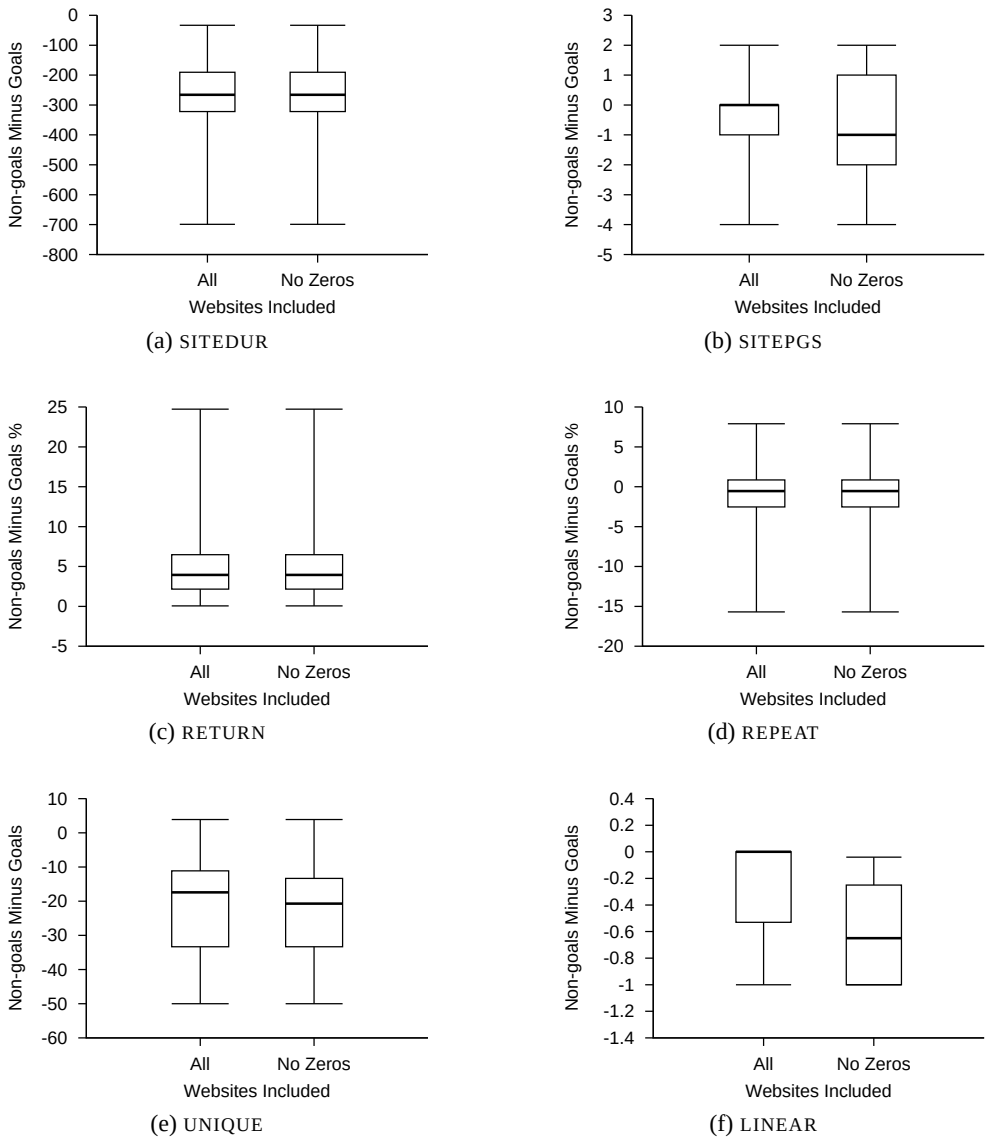
*p ≤ 0.10; **p ≤ 0.05; ***p ≤ 0.01

Table 64: Site-centric: Metric Normality and Skew (Significant – 0.05)

		N		Lilliefors		Shapiro		Skew		Quartile Skew	
Hyp.	Metric	Total	No Zeros	D	p-Value	W	p-Value	All	No Zeros	All	No Zeros
INFORMATION PATCH – PAGE-PATCH											
SC5a	PATCHMAX	14	9	0.19	0.1596	0.88	0.0595*	−0.64	−0.43	−0.15	0.26
SC5b	PATCHLAST	14	9	0.21	0.0966*	0.89	0.0822*	−0.48	−0.56	0.06	−0.53
SC5c	PATCHSUM	14	9	0.21	0.0795*	0.79	0.0033***	−1.41	−1.09	0.11	−0.37
SC6	PATCHDUR	14	14	0.15	0.5374	0.96	0.6601	−0.50	−0.50	−0.02	−0.02
RELAXED INFORMATION SCENT											
SC9a	TRAILMAX	10	6	0.25	0.0671*	0.85	0.0581*	−0.09	0.19	0.08	0.20
SC9b	TRAILLAST	10	6	0.26	0.0616*	0.86	0.0775*	−0.08	−0.31	0.22	0.46
SC9c	TRAILSUM	10	6	0.22	0.1860	0.87	0.0977*	−0.91	−1.05	−0.08	0.07

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

Besides statistics on skew as shown in tables 63 and 64, figures 56 and 57 are also provided to graphically show the distribution of points for each measure.



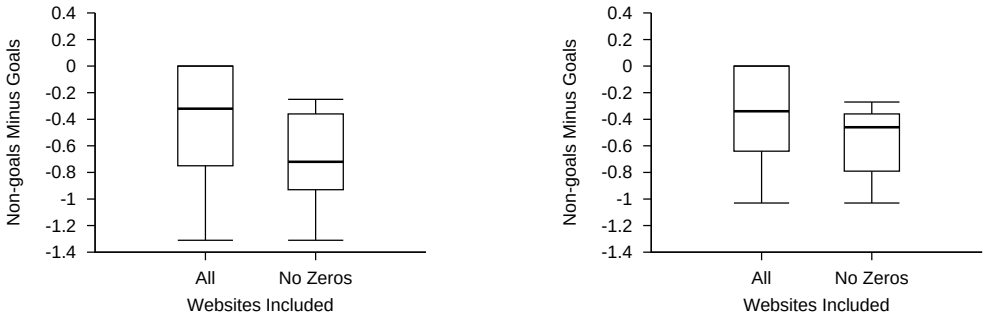
(a) SITEDUR (b) SITEPGS

(c) RETURN (d) REPEAT

(e) UNIQUE (f) LINEAR

Figure 56.: Site-centric: Difference Plots

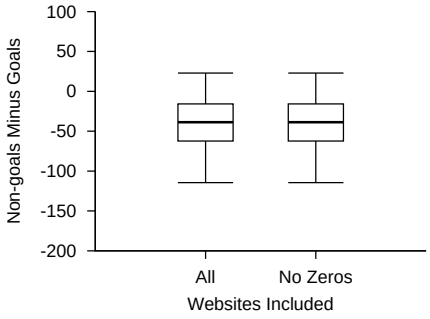
None of the measures exactly met the assumption of no skew (using the quartile skew coefficient). However, PATCHDUR had a very slight negative skew of -0.02 and was considered symmetrical. LINEAR and TRAILSUM also had a slight positive skew values of 0.07 and were considered mostly symmetrical. PATCHLAST had a high amount of skew (-0.53) and did not



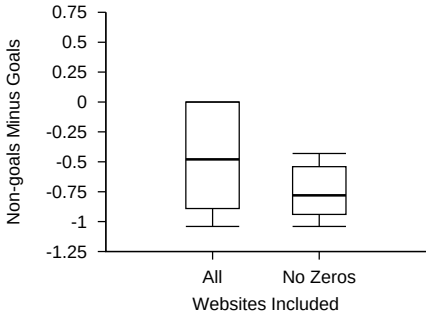
(a) PATCHMAX

(b) PATCHLAST

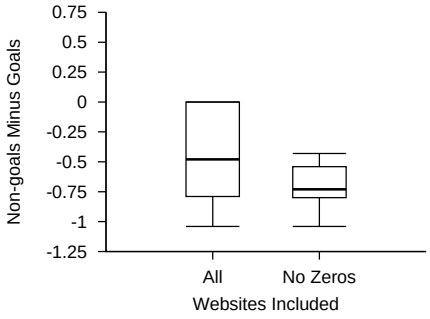
(c) PATCHSUM



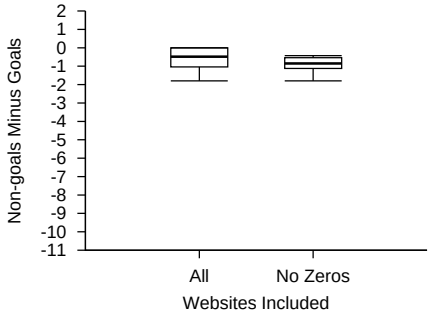
(d) PATCHDUR



(e) TRAILMAX



(f) TRAILLAST



(g) TRAILSUM

Figure 57.: Site-centric: Patch and Trail Difference Plots (Significant – 0.05)

meet the assumption of symmetry. All the other measures did not fully meet the assumption of symmetry since they had slight to moderate amounts of skew ($|0.15|$ to $|0.46|$).

Paired t-Test

Three of the five assumptions for the paired t-test were met. The fourth assumption dealing with symmetry of the D_is was met for some of the metrics, but not for all of the metrics. The fifth assumption of normality was met for only a couple of measures.

Assumption 3 – Each metric used in this research was a quantitative variable measured on at least an interval scale.

Assumption 4 – Each of the differences (D_i) was taken from a Web site from the same population. Therefore, the mean of each difference was expected to be the same.

Assumption 5 – Each of the differences (D_i) was taken from a different and independent Web site.

Assumption 6 – The symmetry for each measure was determined in the same manner as for the exact Wilcoxon signed rank test, except the "All" columns were used to determine skew since the t-test uses points with either a difference or no difference (i.e., both $D_i = 0$ and $D_i \neq 0$) in its calculation. Of the 13 measures, four of the measures had a very slight to slight amount of skew: PATCHLAST (0.06), PATCHDUR (-0.02), TRAILMAX (0.08), and TRAILSUM (-0.08). Two of the measures had a severe skew of -1.00 (SITEPGS and LINEAR) and did not meet the assumption of symmetry. The other seven measures had a slight to moderate amount of skew ($|0.11|$ to $|-0.43|$).

Assumption 7 – Two formal statistical tests of normality were performed on the differences (D_is) for each of the metrics[29]: Lilliefors (Kolmogorov-Smirnov) and Shapiro-Wilk normality tests (Conover, 1999)[30]. Each test has a null hypothesis that the data follows a normal distribution with an unspecified mean and variance. Lilliefors is a non-parametric normality test, whereas

[29] Symmetry is a necessary, but not sufficient, condition for normality. Although SITEPGS and LINEAR were severely skewed and thus not symmetrical, the tests of normality were still performed on the measures for purposes of completeness.

[30] Although presented, formal tests of normality (such as Lilliefors and Shapiro-Wilk) are known to be sensitive to even slight departures from normality (Mendenhall and Sincich, 2003).

Shapiro-Wilk has been found to have greater power than other tests (such as Lilliefors) in many situations (Conover, 1999).

Two of the measures failed to reject the null hypothesis of a normal distribution for both of the Lilliefors and Shapiro-Wilks tests at an α level of 0.15 or lower (UNIQUE and PATCHDUR). PATCHMAX and TRAILSUM also failed to reject the null hypothesis using the Lilliefors test, but did reject the null hypothesis using the more powerful Shapiro-Wilks test at the 0.10 significance level. The remaining nine measures rejected the null hypothesis using both tests with at least an α level of 0.10. Thus, only two of the measures met the assumption of normality, while the distributions of the other 11 metrics were considered non-normal.

In addition to the tests of normality, the skew values from tables 63 and 64 and the graphical depiction of each metric's points (figures 56 and 57) provided further evidence that most of the measures did not follow a normal distribution.

Overall, only the assumptions for the dependent-samples sign test were fully met. Therefore, the sign test was used to test the hypotheses of the site-centric model[31]. The assumptions for the Wilcoxon test and t-test were not completely met and are provided only for comparison purposes. The lack of symmetry (and normality) for some of the measures means the results from the Wilcoxon and t-test must be interpreted with caution.

7.2.2 Hypotheses Testing

Tables 65 and 66 present the results for the nine site-centric hypotheses. Table 65 provides results from the six hypotheses whose measure did not rely on mining patches and trails. Table 66 lists the results for the three hypotheses that required mining patches and trails.

The first two columns of each table list the hypothesis number and name of the metric being tested. The third and fourth columns list the total number of Web sites and the number of sites with a non-zero difference (i.e., $D_i \neq 0$), respectively. The total number of Web sites was used in the t-test, while only Web sites with non-zero differences were used for the Wilcoxon and sign tests. Columns five through seven list the t statistic, degrees of freedom (df), and p-value for the

[31]The sign test is generally the least powerful of the three tests (Conover, 1999). However, as all of the sign test's assumptions were met, greater confidence can be given to the results of the sign test compared to the other two tests.

t-test. The eighth and ninth columns display the V statistic and p-value for the Wilcoxon test. The final two columns list the S statistic and p-value for the sign test[32].

[32]Results of the sign test are presented below since all three assumptions for the test were met. The results of the t-test and Wilcoxon test are provided in footnotes. Since neither the t-test nor the Wilcoxon test met all of their assumptions, the results of those tests should be interpreted with caution.

Table 65: Site-centric: Results

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – SITE-PATCH										
SC1	SITEDUR	47	47	14.85	46	$< 0.0001^{***}$	1,128	$< 0.0001^{***}$	47	$< 0.0001^{***}$
SC2	SITEPGS	47	28	2.27	46	0.0140^{**}	295	0.0166^{**}	19	0.0436^{**}
SC3	RETURN	47	47	−6.66	46	1.0000	0	1.0000	0	1.0000
SC3	(opp)[a]	47	47	6.66	46	$< 0.0001^{***}$	1,128	$< 0.0001^{***}$	47	$< 0.0001^{***}$
SC4	REPEAT	47	47	1.74	46	0.0447^{**}	736	0.0346^{**}	29	0.0719^{*}
STRICT INFORMATION SCENT										
SC7	UNIQUE	47	44	10.19	46	$< 0.0001^{***}$	986	$< 0.0001^{***}$	42	$< 0.0001^{***}$
SC8	LINEAR	47	18	4.41	46	$< 0.0001^{***}$	171	$< 0.0001^{***}$	18	$< 0.0001^{***}$

[a] Hypothesis tested in opposite direction as original – i.e., leaving and returning will be *negatively* associated with achieving a goal on this long tail Web site.
*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

Table 66: Site-centric: Results (Significant – 0.05)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	14	9	3.68	13	0.0014^{***}	45	0.0020^{***}	9	0.0020^{***}
SC5b	PATCHLAST	14	9	3.92	13	0.0009^{***}	45	0.0020^{***}	9	0.0020^{***}
SC5c	PATCHSUM	14	9	3.00	13	0.0051^{**}	45	0.0020^{***}	9	0.0020^{***}
SC6	PATCHDUR	14	14	4.11	13	0.0006^{***}	100	0.0006^{***}	13	0.0009^{***}
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	10	6	3.33	9	0.0044^{**}	21	0.0156^{**}	6	0.0156^{**}
SC9b	TRAILLAST	10	6	3.36	9	0.0042^{**}	21	0.0156^{**}	6	0.0156^{**}
SC9c	TRAILSUM	10	6	2.89	9	0.0089^{**}	21	0.0156^{**}	6	0.0156^{**}

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

SC1 – SITEDUR

Similar to UC1, the first hypothesis of the site-centric model also expected goal achieving foragers would spend more time on a Web site than non-goal achieving foragers[33]. However, the site-centric model did not compare the relative behavior of a forager from one site to another. Instead, all comparisons were done relative to an absolute value of zero. Regardless of how comparisons were determined, the rationale for the hypothesis does not change from one model to another: an expectation of higher durations for goal sessions due to greater information gain from the site of interest.

The results of the sign test supported SC1 at α = 0.01 (S = 47; p-value = < 0.0001)[34]. All 47 Web sites had a higher median duration amongst goal sessions than non-goal sessions. Goal sessions spent roughly six minutes on a site, while non-goal sessions foraged for fewer than two minutes.

The use of *absolute* duration, as used in this hypothesis, provides additional support to prior research regarding the positive association between duration and goal achievement. In addition, although the results between the user- and site-centric models are not directly comparable due to the relative nature of the user-centric model, the significant support for both hypotheses generally reinforces one another regarding the value of duration to predict goal achievement.

SC2 – SITEPGS

Both the first and second hypotheses were similar to one another because they both focused on the concept of satisficing (Pirolli, 2007; Simon, 1956). However, SC2 examined the importance of greater pages viewed at a Web site rather than duration. The belief was that every additional page signaled the continued interest of the forager to stay on the site because information of value could still be obtained from the Web site. Thus, greater pages equated to more information of value being obtained which further lead to a greater probability of achieving a goal by the forager.

The second hypothesis was found to be significant at α = 0.05 (S = 19; p-value = 0.0436)[35],

[33] The rationale for the first four site-centric hypothesis are explained in greater detail in the user-centric results (§7.1.2).

[34] Hypothesis SC1 was also significant at α = 0.01 for both the t-test (t = 14.85; df = 46; p-value = < 0.0001) and Wilcoxon test (V = 1,128; p-value = < 0.0001).

[35] Hypothesis SC2 was also significant at α = 0.05 for both the t-test (t = 2.27; df = 46; p-value = 0.0140) and Wilcoxon test (V = 295; p-value = 0.0166).

supporting hypothesis SC2. 19 out of the 28 non-tied Web sites had a higher median number of pages viewed for goal sessions versus non-goal sessions. Goal sessions viewed 4.28 pages on average, whereas non-goal sessions viewed only 3.85 pages.

Hypothesis SC2 provides additional support to prior literature about the positive association between number of pages viewed and goal achievement. When compared to the user-centric model, however; the results for this hypothesis were less significant. One possible reason for the difference may be due to the use of absolute rather than relative comparisons of foragers' behavior. However, another likely reason may be due to the structure of the Web sites used in the site-centric dataset.

In general, site-centric Web sites had relatively few pages (16.36 pages on average). While the number of pages on the user-centric Web sites was unknown, the sites may have had more pages than the site-centric Web sites[36]. Therefore, there would be more pages that could have been visited on a Web site from the user-centric dataset, leading to a greater gap between page visitation behavior of goal and non-goal sessions, and hence a more significant result.

SC3 – RETURN

Hypothesis SC3 conjectured that foragers who left a Web site and returned during the same session would be more likely to achieve a goal. The hypothesis was not supported at any of the tested α levels (S = 0; p-value = 1.0000)[37]. None of the 47 Web sites had a higher percentage of goals achieved amongst foragers who left the site and returned during the same session than visitors that stayed on the site during their entire session. The results indicated a forager was *less* likely to achieve a goal if the user left a Web site and returned within the same session (the opposite of hypothesis SC3), which was significant at α = 0.01 (S = 47; p-value = < 0.0001)[38]. All 47 Web sites found a higher proportion of goal sessions that stayed rather than left and returned during their session.

[36] On average, foragers from the Web sites in the user-centric dataset viewed 14.84 pages during a session, while visitors to site-centric sites only viewed 4.06 pages. While not definitive proof that user-centric sites had more pages than site-centric Web sites, the higher average number of pages viewed for user-centric foragers does give some indication that those Web sites might have had more pages.

[37] Hypothesis SC3 was also not supported at any of the tested α levels for both the t-test (t = -6.66; df = 46; p-value = 1.0000) and Wilcoxon test (V = 0; p-value = 1.0000).

[38] The opposite of hypothesis SC3 was also supported at α = 0.01 for both the t-test (t = 6.66; df = 46; p-value = < 0.0001) and Wilcoxon test (V = 1,128; p-value = < 0.0001).

Since prior research has not examined returning behavior of a user during the same session, this hypothesis (in opposite form) provides an initial result of a negative association between returning behavior during the same session and goal achievement. In comparison to the user-centric model, the result of this hypothesis was also not supported. However, unlike the user-centric model, the opposite of the original hypothesis was supported at α = 0.01. The difference in support and not support between the two models may have been due to the methodology of determining when leaving and returning behavior occurred.

In general, the user-centric model was conservative while the site-centric model was liberal when classifying leaving and returning behavior. For example, the user-centric model only counted visits of at least two pages to other known e-commerce Web sites as valid leaving behavior. Such a precise manner of determining leaving behavior was not possible in the site-centric data. Therefore, a more simplistic manner of determining leaving behavior was used which examined the referring field for each page viewed. A limitation of using the referring field was it was not known if true browsing behavior (i.e., more than one page was viewed) took place at the referred Web site. Thus situations in which a forager left the site of interest, viewed one page on another site, and then returned would still be marked as leaving and returning in the site-centric model[39]. In addition, the referring field was also limited because it was unable to capture if a forager left to view another Web site in a new Web browser window or tab.

SC4 – REPEAT

The final hypothesis which examined the value of the entire Web site as a patch looked at how prior visitation behavior would affect goal achievement. Hypothesis SC4 expected prior visitation of a Web site would provide a forager with intimate knowledge of what the site had to offer. Therefore, when the forager has some need to be met at a future date, they would more likely return to the Web site of interest if they believed it would satisfy their information goal. Thus, repeat visitation would signal greater likelihood of achieving a goal at the Web site of interest.

The results of the sign test demonstrated the fourth hypothesis was significant at α = 0.10 (S = 29; p-value = 0.0719)[40], supporting hypothesis SC4. 29 of the 47 Web sites had a higher median

[39] Another example would be a forager that clicked the back button of their browser one too many times and ended up on the search engine page that initially brought them to the site of interest, and then clicked a link to return back to the site of interest.

[40] Hypothesis SC4 was significant at α = 0.05 for both the t-test (t = 1.74; df = 46; p-value = 0.0447) and Wilcoxon

probability of a form submission amongst goal sessions when a user had visited the site before. The slightly significant result was due to the small difference between the proportion of goal sessions that had and had not visited before (0.62% difference between groups).

Prior literature has found mixed associations (dependent on the task) between users returning during different sessions and completing a task (Sismeiro and Bucklin, 2004). The result of this hypothesis lends additional support of a slight positive association between repeat visitation behavior and goal achievement. When compared to the user-centric model, there was a large difference in significance between the two models (α = 0.01 user-centric versus 0.10 site-centric). The difference in significance may be partially explained in two ways.

First, the nature of the goal being examined in the user-centric dataset may better lend itself to repeat visitation than the goal in the site-centric dataset. For example, the user-centric dataset examined product purchases, which may need to be replenished from time to time. In contrast, there is likely little need to resubmit contact information on one of the site-centric Web sites. From another standpoint, a purchase has a defined cost associated with it. Therefore, a forager may return to a site multiple times as they contemplate purchasing a product. Leaving contact information on a Web site has no real monetary cost associated with the action. Therefore, the submission of a contact form may not require the same degree of thought and comparison that purchasing does, which may lower the need for repeat visitations to a site.

The second reason a difference between the hypotheses of the two models was seen may be due to the mechanism by which site-centric foragers were identified in the dataset. Cookies were used to identify and track users across sessions. If a user deleted their cookie then they would be seen as a new visitor on any subsequent visit. Thus, repeat visitation of foragers may be underrepresented in the site-centric dataset.

SC5 – PATCHMAX, PATCHLAST, and PATCHSUM

The fifth hypothesis expected that visitation of goal patches would be positively associated with goal achievement. The hypothesis operated under the assumption that certain areas of a Web site were more valuable to goal achieving foragers than other areas of the site. Thus, users that visited those same areas of the Web site were assumed to have similar information goals and should be

test (V = 736; p-value = 0.0346). The reason the t-test and Wilcoxon test found SC4 significant at a lower alpha level than the sign test may be due to the lower power of the sign test.

more likely to achieve a goal. The actual value from a forager's visitation of patches was specified in slightly different ways in three sub-hypotheses: maximum value of a patch, value of last patch visited, and total value of all patches visited.

The three sub-hypotheses of SC5 were all found to be significant at α = 0.01 (S = 9; p-value = 0.0020 for all three measures)[41], supporting hypotheses SC5a-c (table 66). 14 Web sites out of the 47 total Web sites discovered patches at the 0.05 significance level, with only nine of those 14 sites having a non-zero difference. All nine of the non-zero Web sites had goal sessions with higher median values for the most valuable patch visited, last patch visited, and sum of all patches visited.

On average, foragers visited 2.75 patches per session. With so few patches being visited it was possible that some of the measures did not differ from one another by a great deal. For example, sessions that only visited a single patch would have the same value for all three measures. However, as foragers visited almost three patches per session, the average value of PATCHSUM was at least 0.42 points higher than either of the other measures, making it unlikely the PATCHSUM measure only included the same patches as the other two measures. For the other two measures though, the average difference between PATCHMAX and PATCHLAST was only 0.03 points, indicating many sessions may have had the same patch be the most valuable and last patch visited. Therefore, even though both sub-hypotheses were supported, the similarity of each measure means the actual impact of the most valuable and last visited patch on goal achievement cannot be reliably separated from one another.

Since prior research has not examined the impact groups of pages (that may be of different types (e.g., product pages, informational pages)), this hypothesis provides an initial result of a positive association between visitation of patches and goal achievement.

SC6 – PATCHDUR

The sixth hypothesis expected that mere visitation alone of valuable patches would not necessarily mean foragers were obtaining value from those patches. Therefore, similar to hypothesis SC1, this hypothesis also relied on the concept of satisficing (Pirolli, 2007; Simon, 1956); contending

[41] Using the t-test, hypotheses SC5a and SC5b were both significant at α = 0.01 (PATCHMAX (t = 3.68; df = 13; p-value = 0.0014); PATCHLAST (t = 3.92; df = 13; p-value = 0.0009)), while SC5c was significant at α = 0.05 (t = 3.00; df = 13; p-value = 0.0051). Using the Wilcoxon test, all three hypotheses were significant at α = 0.01 (V = 45; p-value = 0.0020 for all three measures). The discrepancy between the significance of PATCHSUM from the t-test versus the Wilcoxon and sign test may be due to a lack of normality of the measure. Without normality, the t-test may not have enough power to detect as significant of a difference as the other two tests.

higher amounts of time spent within patches related to more information gained and thus a greater likelihood of goal achievement.

Hypothesis SC6 found a higher median duration within patches for goal sessions than non-goal sessions at α = 0.01 (S = 13; p-value = 0.0009)[42], supporting the hypothesis. 13 of the 14 non-zero Web sites with discovered patches had goal sessions spend a higher median duration of time within patches than non-goal sessions spent in patches. On average, goal sessions spent almost three-quarters of a minute more in patches than non-goal sessions.

The use of duration on an entire site has been used in prior literature to explain choice behavior, with mixed results. Duration has had mixed associations with purchasing for different tasks (Sismeiro and Bucklin, 2004) along with differences in significance among different datasets (Padmanabhan et al., 2001). However, as prior literature has not examined the concept of patches before, this hypothesis provides an initial result of a positive association between duration within patches and goal achievement.

SC7 – UNIQUE

Hypothesis SC7 was the first of two hypotheses that defined information scent in a strict manner. In addition, the hypothesis also viewed a forager's session as a single monolithic piece, and assumed any repeat page viewings (regardless of location) were indicative of poor scent. In turn, the cause of lackluster information scent was believed to be either a poorly defined information goal or a less than optimal Web site design, both of which were less likely to result in a goal being achieved. Stated in a positive direction, the hypothesis proposed that the lower the percentage of duplicate pages viewed, the more likely a goal would be achieved.

Hypothesis SC7 (table 65) was significant at α = 0.01 (S = 42; p-value = $<$ 0.0001)[43], supporting the hypothesis that goal achieving sessions would visit fewer duplicate pages than non-goal sessions. 42 of the 44 non-zero Web sites had goal sessions with a higher median percentage of unique pages viewed than non-goal sessions, with goal sessions viewing almost 90$\%$ unique pages and non-goal sessions only viewing about 70$\%$.

Prior research has examined the relationship between the proportion of unique pages visited and

[42]Hypothesis SC6 was also significant at α = 0.01 for both the t-test (t = 4.11; df = 13; p-value = 0.0006) and Wilcoxon test (V = 100; p-value = 0.0006).

[43]Hypothesis SC7 was also significant at α = 0.01 for both the t-test (t = 10.19; df = 46; p-value = $<$ 0.0001) and Wilcoxon test (V = 986; p-value = $<$ 0.0001).

purchasing behavior. Moe (2003) found that the proportion of unique pages differed depending on the type of pages being viewed (e.g., brand pages, product pages, category pages). However, this hypothesis examined proportion of unique pages across all page types. Therefore, hypothesis SC7 provides support for the general positive association between proportion of unique pages viewed and goal achievement.

SC8 – LINEAR

Hypothesis SC8 was the second of the two hypotheses that defined information scent in a strict manner, where repeat visitations were viewed as indications of poor scent. However, this hypothesis took a finer-grained conceptualization of scent than the previous hypothesis by examining the complexity of a user's session. Complexity was determined by not only what pages were viewed, but also the order in which they were viewed. Hypothesis SC8 proposed that less complex (i.e., more linear) clickstreams were indicative of higher levels of scent, and thus a greater likelihood of achieving a goal.

The hypothesis was found to be significant at α = 0.01 (S = 18; p-value = < 0.0001)[44], supporting hypothesis SC8. All 18 of the non-zero Web sites had higher median linear clickstream values for goal sessions compared to non-goal sessions. The average goal sessions had a linear clickstream value of 0.98 compared to the average value of 0.75 for non-goal sessions.

Prior research has found success in using the measure of session complexity to distinguish between groups (McEneaney, 2001), in the use of product recommendation agents (Senecal et al., 2005), and in predicting the completion of information and e-commerce tasks (Kalczynski et al., 2006). This hypothesis strengthens the use of session complexity to distinguish between goal and non-goal sessions within the context of goal achievement.

Both of the two strict information scent hypotheses were found to be supported at the same α level. Using the results of the sign test, one measure was not able to be definitively considered better than the other, since the number of non-zero Web sites was different for each hypothesis[45]. In addition, even though the LINEAR measure had a greater percentage of its non-zero Web sites in

[44] Hypothesis SC8 was also significant at α = 0.01 for both the t-test (t = 4.41; df = 46; p-value = < 0.0001) and Wilcoxon test (V = 171; p-value = < 0.0001).

[45] Examining the t-test shows a clear preference for the UNIQUE measure in being better able to distinguish between goal and non-goal sessions (t value of 10.19 versus 4.41). However, as the assumptions of the t-test were not fully met, the results of the t-test should be interpreted with caution.

the positive direction (100% versus 95.45%), the loss of two Web sites in the negative direction for the UNIQUE measure was not enough to raise the p-value precipitously.

SC9 – TRAILMAX, TRAILLAST, and TRAILSUM

The final hypothesis examined information scent from a relaxed viewpoint. The hypothesis assumed that foragers who followed trails predominately traversed by prior goal sessions would be positively associated with goal achievement. The hypothesis operated under the assumption that certain paths throughout a Web site (with or without "inefficiencies") were indicators of high scent relevant to a goal-achieving information goal. The value obtained from a forager following a trail was specified in three slightly different sub-hypotheses: maximum value of a trail, value of last trail followed, and total value of all trails followed.

The three sub-hypotheses of SC9 (table 66) regarding the following of valuable trails as a means to explain goal achievement were all found to be significant at α = 0.05 (S = 6; p-value = 0.0156 for all three measures)[46], supporting hypotheses SC9a-c. 10 Web sites out of the 47 total sites discovered trails at the 0.05 significance level, with only six of those 10 sites having a non-zero difference. All six of the non-zero Web sites had goal sessions with higher median values for the most valuable trail followed, last trail followed, and sum of all trails followed.

On average, foragers followed 1.60 trails per session. As many foragers only followed a single trail per session, it was likely the measures did not differ from one another by a great deal. For example, sessions that only followed a single trail would have the same value for all three measures. Examining the difference in value between the three measures (0.26 to 0.33) failed to reveal a clear and distinct difference between them. Therefore, even though all three sub-hypotheses were supported, the similarity of each measure means the actual impact of the most valuable, last followed, and total value of all trails followed on goal achievement cannot be reliably separated from one another.

Prior research has examined the use of paths and portions of paths to predict future patch selections (Montgomery et al., 2004; Yang et al., 2004). However, the use of path fragments to segment groups of a Web site population has not been examined in prior literature. Thus, this hypothesis

[46]Hypotheses SC9a-c were significant at α = 0.05 for both the t-test (TRAILMAX (t = 3.33; df = 9; p-value = 0.0044); TRAILLAST (t = 3.36; df = 9; p-value = 0.0042); TRAILSUM (t = 2.89; df = 9; p-value = 0.0089)) and Wilcoxon test (V = 21; p-value = 0.0156 for all three measures).

provides an initial result of a positive association between following of trails and goal achievement.

Summary of Results

Table 67 summarizes the results of the hypotheses testing. Of the 13 hypotheses and sub-hypotheses, seven were supported at α = 0.01, four at α = 0.05, and one at α = 0.10. Hypothesis SC3 was not supported in its original form, but the opposite of SC3 was supported at α = 0.01.

Table 67: Site-centric: Hypotheses Results Summary

Hyp.	Metric	Hypothesis Supported?
INFORMATION PATCH – SITE-PATCH		
SC1	SITEDUR	Yes***
SC2	SITEPGS	Yes**
SC3	RETURN	No
SC3	(opp)[a]	Yes***
SC4	REPEAT	Yes*
INFORMATION PATCH – PAGE-PATCH		
SC5a	PATCHMAX	Yes***
SC5b	PATCHLAST	Yes***
SC5c	PATCHSUM	Yes***
SC6	PATCHDUR	Yes***
RELAXED INFORMATION SCENT		
SC7	UNIQUE	Yes***
SC8	LINEAR	Yes***
RELAXED INFORMATION SCENT		
SC9a	TRAILMAX	Yes**
SC9b	TRAILLAST	Yes**
SC9c	TRAILSUM	Yes**

[a] Hypothesis tested in opposite direction as original – i.e., leaving and returning will be *negatively* associated with achieving a goal on this long tail Web site.
*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

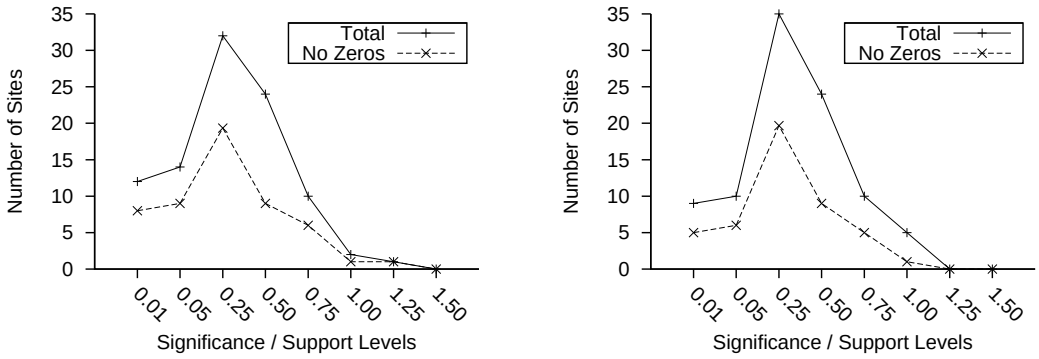
7.2.3 Sensitivity Analysis

The previous section tested hypotheses SC5a-c, SC6, and SC9a-c from patches and trails mined at the 0.05 significance level. The use of α = 0.05 for learning patches and trails was motivated by prior research that used the same α level when discovering contrast sets (Bay and Pazzani, 1999). However, other significance levels and different means of detecting contrast sets may be used (e.g., support). Therefore, a sensitivity analysis was done to see how the selection of mining criteria used for learning patches and trails may affect the results of the hypotheses.

This section provides descriptive statistics and results for hypotheses SC5a-c, SC6, and SC9a-c at two different significance levels (0.01 and 0.05) and six distinct support levels (0.25 – 1.50 in 0.25 increments).

Descriptive Statistics

Figure 58 illustrates the number of Web sites that discovered patches (figure 58a) and trails (figure 58b) from all eight mined significance and support levels used[47]. Each figure also displays the number of Web sites which did not have a zero difference for the tested measures[48].



(a) Patches (b) Trails

Figure 58.: Site-centric: Patch and Trail Sample Size by Significance / Support Levels

Between the two mined significance levels, there was very little change in the number of Web sites for either patches or trails. An increase of only two Web sites for patches (16.67%) and one

[47]The actual number of Web sites may be found in tables 69 – 76, which are introduced later in this subsection.

[48]The number of non-zero Web sites was determined by finding the average number of "no zero" Web sites from hypotheses SC5a-c for patches and SC9a-c for trails.

Web site for trails (11.11%) was seen when moving from the more stringent α = 0.01 to the less stringent α = 0.05.

The difference in sample size between mined support levels was much more dramatic than between mined significance levels. Higher support levels greatly reduced the number of Web sites which discovered patches and trails of the specified value. At the 0.25 support level, 32 Web sites found patches and 35 sites discovered trails. An increase to the 0.50 support level saw a 25.00% drop in patch Web sites (to 24 sites) and a 31.43% decrease in trail Web sites (to 24 sites). An even greater percentage drop in Web sites was seen using the 0.75 support level: 58.33% decrease in Web sites for both patches and trails (to 10 sites). At the higher support levels there were very few Web sites discovering patches or trails. Only two and five Web sites, and one and zero Web sites for patches and trails were found at the 1.00 and 1.25 support levels, respectively.

Figure 59 displays the average value of all sessions for the three patch visitation (figure 59a) and trail following (figure 59b) measures across all eight mining significance and support levels used[49]. In addition, figure 59 also shows the average number of seconds spent within patches for all sessions (figure 59c).

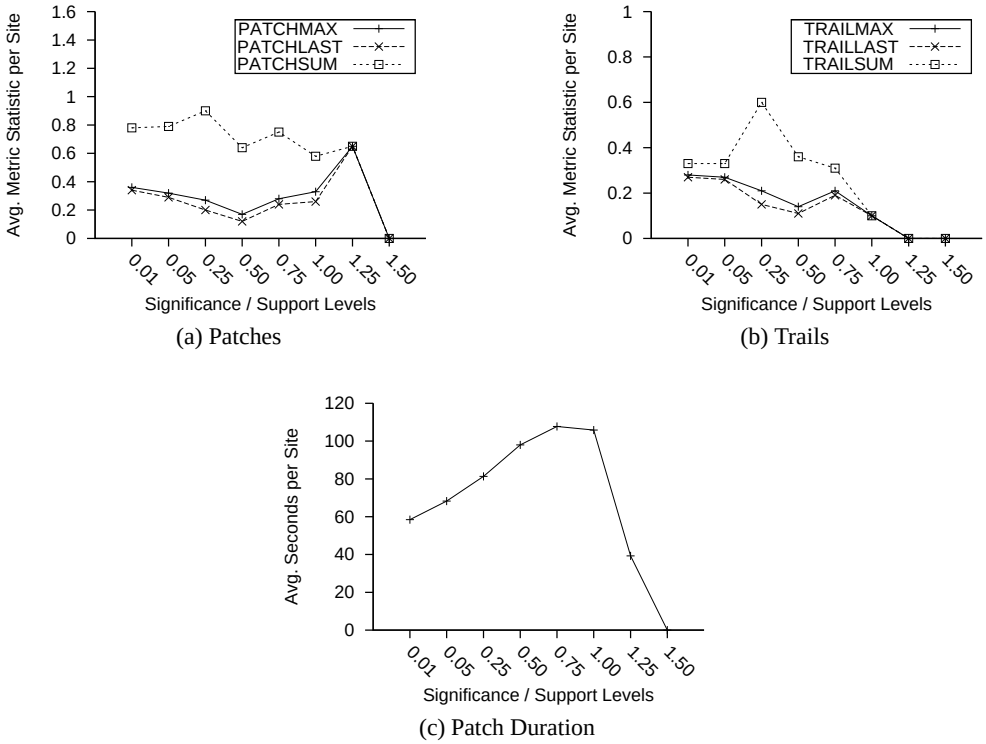
In general, the average values for each of the measures appeared to stay within a relatively narrow range of one another from the 0.01 significance level to the 0.75 support level. Table 68 further reinforces the relative stability of these measures by listing the mean, standard deviation, median, minimum, and maximum values from the average of the first six significance and support levels. The standard deviation for the three patch visitation and three trail following measures ranged from 0.07 to 0.13.

The highest values of PATCHMAX and PATCHLAST were found at the 0.01 significance level (0.36 and 0.34), while TRAILMAX and TRAILLAST were at their highest average values at the 0.05 significance level (0.37 and 0.35). Not surprisingly, both of the sum measures (PATCHSUM and TRAILSUM) had their highest values (0.90 and 0.60) when the support was 0.25 (when the most number of patches and trails were discovered[50]).

The PATCHDUR measure was the only metric that continued to increase through all the signifi-

[49]The results from the 1.00 support level and above should be interpreted with caution as the averages were calculated from very few Web sites. In addition, the averages displayed in the figures were calculated by including sessions which did not visit a patch or trail. Therefore, the average metric may be lower than should otherwise be possible. For example, the lowest average of patches learned at the 0.50 support level should be 0.50. However, the average PATCHMAX value was 0.17 for all sessions.

[50]See tables 77 and 78 for more details.



(a) Patches

(b) Trails

(c) Patch Duration

Figure 59.: Site-centric: All Patch and Trail Metrics by Significance / Support Levels

cance and support mining levels[51]. The increase may have been due to three inter-related reasons.

First, the average number of patches per site increased from the 0.01 significance level to the 0.25 support level. An increased number of patches generally meant greater coverage of the Web site (e.g., 23.37% to 42.22% coverage between the 0.01 significance level and 0.25 support level)[52]. Therefore, the duration spent in patches may have more closely aligned with the amount of time a forager spent on the Web site as a whole.

The second reason may be due to the increase in average patch size. For example, the average size of patches went from 1.67 to 2.25 pages when going from the 0.01 significance level to the 0.50 support level. When the size of a patch was increased then the total duration within the patch included the duration of more pages. Therefore, an increased total duration within a patch may then lead to higher median patch durations.

The final reason may have been because the average value of a patch increased. For example, between the 0.25 to 0.75 support levels the average patch value increased from 0.42 to 0.88. The assumption was foragers would spend more time within the more valuable patches. Therefore, when a site only had valuable patches (e.g., at support level 0.75), then (1) more time should have been spent within those patches and (2) the median patch duration of the forager was not reduced by the visitation of less valuable patches (where less time within the patch would be expected).

Table 68: Site-centric: Sensitivity Analysis Metric Statistics

	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH					
PATCHMAX	0.28	0.07	0.28	0.17	0.36
PATCHLAST	0.24	0.08	0.24	0.12	0.34
PATCHSUM	0.77	0.09	0.78	0.64	0.90
PATCHDUR (in seconds)	82.75	20.35	81.35	58.46	107.71
RELAXED INFORMATION SCENT					
TRAILMAX	0.24	0.09	0.21	0.14	0.37
TRAILLAST	0.21	0.10	0.19	0.11	0.35
TRAILSUM	0.42	0.13	0.36	0.31	0.60

[51]The measure was calculated by finding the median amount of time spent in all patches by a forager.
[52]Statistics about patch characteristics may be found in § 7.2.3.

Figure 60 expands upon figure 59 by illustrating the average value of the seven measures against three groups of sessions: all, goal, and non-goal[53].

Tables 69 – 76 list the mean, standard deviation, median, minimum, and maximum values for the seven metrics that were calculated from mined patches and trails at the eight different significance and support levels. The statistics for the metrics are displayed in three groups of sessions: all, goal, and non-goal.

Patch and Trail Descriptive Statistics

Figure 61 illustrates the differences between the significance and support levels for four different statistics[54]: number of patches and trails (figure 61a), size of patches and trails (figure 61b), percentage of coverage of patches and trails (figure 61c), and the value of patches and trails (figure 61d). Each figure displays the statistic of patches and trails for each of the metrics. Figure 61 also displays the number of patches visited (figure 61e) and trails followed (figure 61f) from three groups of sessions: all, goal, and non-goal.

Tables 77 – 86 list the mean, standard deviation, median, minimum, and maximum values for the metrics displayed in figure 61 across all eight significance and support mining levels.

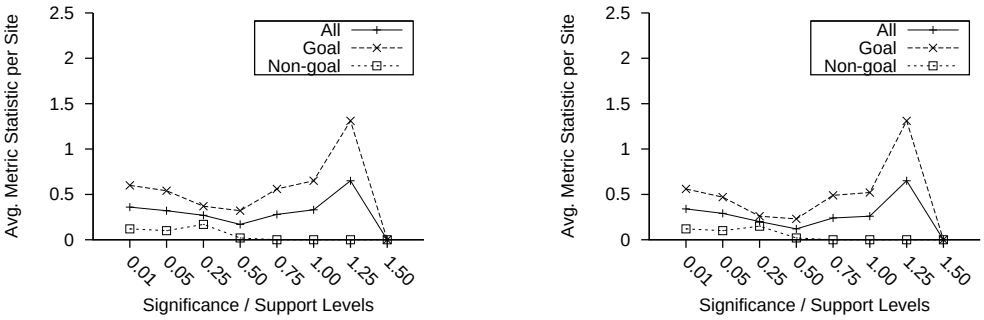
The average number of patches and trails found on a Web site followed a similar pattern for the first five levels, with more patches being discovered than trails. Not surprisingly, the greatest number of average patches (50.16) and trails (39.60) were found using the least stringent support level (0.25). In comparing the significance and support levels, there was not a direct equivalent of either significance level found within the selected support levels. For example, in order to obtain a similar number of patches and trails as found at α = 0.05 (11.93 patches and 4.70 trails), the support level would need to have been between 0.75 and 1.00 (8.00 – 20.20 patches and 1.60 – 7.20 trails).

The average size of patches and trails roughly followed a $\cap$ shape over the significance and support levels, with trails being larger in size than patches for all but the 0.50 support level (2.63 pages per patch versus 2.56 pages per trail)[55]. Patches and trails discovered using significance were smaller in size than those patches and trails found from the first three support levels. For example, patches were 1.82 pages in size and trails were 2.15 pages long at α = 0.05. The first

[53]The actual values used in the figures may be found in tables 69 – 76, which are introduced later in this subsection.

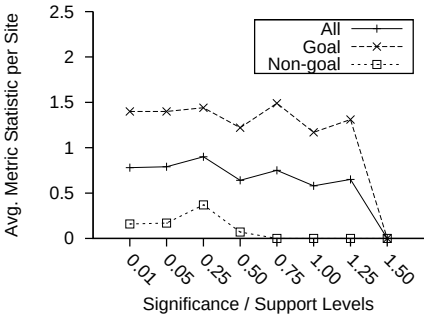
[54]The analysis of patch and trail descriptive statistics do not include support levels greater than 0.75, as a limited number of Web sites found patches and trails at those support levels to provide reliable metric averages.

[55]Trails were restricted to a minimum of two pages in sequence, whereas patches could be one page in size.

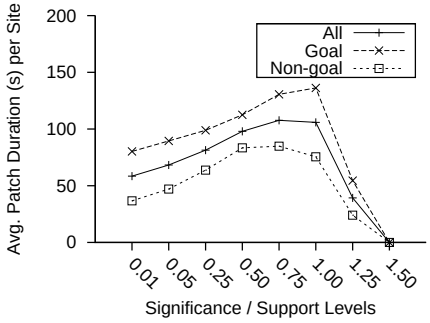


(a) PATCHMAX

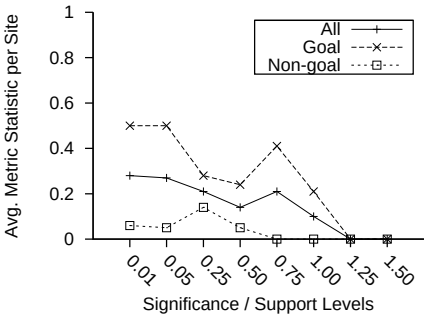
(b) PATCHLAST



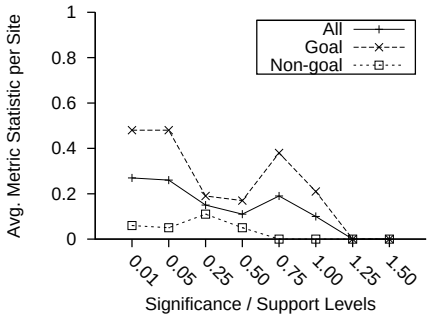
(c) PATCHSUM



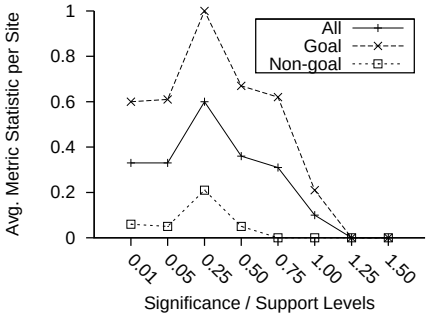
(d) PATCHDUR



(e) TRAILMAX



(f) TRAILLAST



(g) TRAILSUM

Figure 60.: Site-centric: Patch and Trail Metrics by Significance / Support Levels

Table 69: Site-centric: Metric Statistics (Significant – 0.01)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	12	0.36	0.41	0.20	0.00	1.31
Goal		0.60	0.43	0.74	0.00	1.31
Non-goal		0.12	0.21	0.00	0.00	0.53
PATCHLAST						
All	12	0.34	0.37	0.20	0.00	1.03
Goal		0.56	0.37	0.70	0.00	1.03
Non-goal		0.12	0.21	0.00	0.00	0.53
PATCHSUM						
All	12	0.78	1.18	0.20	0.00	4.53
Goal		1.40	1.41	1.06	0.00	4.53
Non-goal		0.16	0.31	0.00	0.00	0.95
PATCHDUR (in seconds)						
All	12	58.46	40.41	47.75	13.00	162.75
Goal		80.23	46.75	71.25	17.00	162.75
Non-goal		36.69	13.97	37.88	13.00	62.50
RELAXED INFORMATION SCENT						
TRAILMAX						
All	9	0.28	0.38	0.00	0.00	1.04
Goal		0.50	0.42	0.50	0.00	1.04
Non-goal		0.06	0.17	0.00	0.00	0.50
TRAILLAST						
All	9	0.27	0.37	0.00	0.00	1.04
Goal		0.48	0.40	0.50	0.00	1.04
Non-goal		0.06	0.17	0.00	0.00	0.50
TRAILSUM						
All	9	0.33	0.51	0.00	0.00	1.80
Goal		0.60	0.60	0.50	0.00	1.80
Non-goal		0.06	0.17	0.00	0.00	0.50

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 70: Site-centric: Metric Statistics (Significant – 0.05)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	14	0.32	0.40	0.00	0.00	1.31
Goal		0.54	0.43	0.60	0.00	1.31
Non-goal		0.10	0.21	0.00	0.00	0.57
PATCHLAST						
All	14	0.29	0.34	0.00	0.00	1.03
Goal		0.47	0.36	0.47	0.00	1.03
Non-goal		0.10	0.20	0.00	0.00	0.51
PATCHSUM						
All	14	0.79	1.25	0.00	0.00	4.53
Goal		1.40	1.52	1.06	0.00	4.53
Non-goal		0.17	0.36	0.00	0.00	1.04
PATCHDUR (in seconds)						
All	14	68.28	47.02	51.38	19.00	178.00
Goal		89.48	53.06	76.63	19.00	178.00
Non-goal		47.07	28.43	38.88	20.00	134.00
RELAXED INFORMATION SCENT						
TRAILMAX						
All	10	0.27	0.37	0.00	0.00	1.04
Goal		0.50	0.40	0.52	0.00	1.04
Non-goal		0.05	0.16	0.00	0.00	0.50
TRAILLAST						
All	10	0.26	0.35	0.00	0.00	1.04
Goal		0.48	0.37	0.52	0.00	1.04
Non-goal		0.05	0.16	0.00	0.00	0.50
TRAILSUM						
All	10	0.33	0.50	0.00	0.00	1.80
Goal		0.61	0.58	0.52	0.00	1.80
Non-goal		0.05	0.16	0.00	0.00	0.50

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 71: Site-centric: Metric Statistics (Supported – 0.25)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	32	0.27	0.31	0.26	0.00	1.31
Goal		0.37	0.36	0.32	0.00	1.31
Non-goal		0.17	0.21	0.00	0.00	0.57
PATCHLAST						
All	32	0.20	0.20	0.26	0.00	0.68
Goal		0.26	0.22	0.28	0.00	0.68
Non-goal		0.15	0.18	0.00	0.00	0.48
PATCHSUM						
All	32	0.90	1.68	0.27	0.00	6.99
Goal		1.44	2.20	0.60	0.00	6.99
Non-goal		0.37	0.56	0.00	0.00	2.05
PATCHDUR (in seconds)						
All	31	81.35	51.13	71.13	16.50	274.00
Goal		98.89	59.29	89.00	26.00	274.00
Non-goal		63.82	34.13	51.75	16.50	130.00
RELAXED INFORMATION SCENT						
TRAILMAX						
All	35	0.21	0.27	0.00	0.00	1.04
Goal		0.28	0.32	0.25	0.00	1.04
Non-goal		0.14	0.20	0.00	0.00	0.58
TRAILLAST						
All	35	0.15	0.19	0.00	0.00	0.63
Goal		0.19	0.20	0.25	0.00	0.63
Non-goal		0.11	0.16	0.00	0.00	0.50
TRAILSUM						
All	35	0.60	1.57	0.00	0.00	10.72
Goal		1.00	2.14	0.25	0.00	10.72
Non-goal		0.21	0.33	0.00	0.00	1.13

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 72: Site-centric: Metric Statistics (Supported – 0.50)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	24	0.17	0.35	0.00	0.00	1.31
Goal		0.32	0.44	0.00	0.00	1.31
Non-goal		0.02	0.12	0.00	0.00	0.57
PATCHLAST						
All	24	0.12	0.25	0.00	0.00	0.79
Goal		0.23	0.31	0.00	0.00	0.79
Non-goal		0.02	0.10	0.00	0.00	0.51
PATCHSUM						
All	24	0.64	1.62	0.00	0.00	6.35
Goal		1.22	2.14	0.00	0.00	6.35
Non-goal		0.07	0.32	0.00	0.00	1.58
PATCHDUR (in seconds)						
All	24	97.97	60.06	84.00	13.00	348.25
Goal		112.55	67.35	95.25	17.00	348.25
Non-goal		83.40	48.90	70.75	13.00	171.75
RELAXED INFORMATION SCENT						
TRAILMAX						
All	24	0.14	0.30	0.00	0.00	1.04
Goal		0.24	0.37	0.00	0.00	1.04
Non-goal		0.05	0.16	0.00	0.00	0.58
TRAILLAST						
All	24	0.11	0.22	0.00	0.00	0.68
Goal		0.17	0.25	0.00	0.00	0.68
Non-goal		0.05	0.16	0.00	0.00	0.58
TRAILSUM						
All	24	0.36	1.01	0.00	0.00	5.55
Goal		0.67	1.35	0.00	0.00	5.55
Non-goal		0.05	0.16	0.00	0.00	0.58

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 73: Site-centric: Metric Statistics (Supported – 0.75)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	10	0.28	0.45	0.00	0.00	1.31
Goal		0.56	0.50	0.76	0.00	1.31
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHLAST						
All	10	0.24	0.39	0.00	0.00	1.03
Goal		0.49	0.43	0.76	0.00	1.03
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHSUM						
All	10	0.75	1.43	0.00	0.00	4.33
Goal		1.49	1.75	0.76	0.00	4.33
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHDUR (in seconds)						
All	10	107.71	90.64	89.38	13.00	398.25
Goal		130.60	109.67	105.88	17.00	398.25
Non-goal		84.83	64.42	71.00	13.00	240.75
RELAXED INFORMATION SCENT						
TRAILMAX						
All	10	0.21	0.38	0.00	0.00	1.04
Goal		0.41	0.46	0.22	0.00	1.04
Non-goal		0.00	0.00	0.00	0.00	0.00
TRAILLAST						
All	10	0.19	0.36	0.00	0.00	1.04
Goal		0.38	0.43	0.20	0.00	1.04
Non-goal		0.00	0.00	0.00	0.00	0.00
TRAILSUM						
All	10	0.31	0.61	0.00	0.00	1.81
Goal		0.62	0.76	0.22	0.00	1.81
Non-goal		0.00	0.00	0.00	0.00	0.00

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 74: Site-centric: Metric Statistics (Significant – 1.00)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	2	0.33	0.65	0.00	0.00	1.31
Goal		0.65	0.92	0.65	0.00	1.31
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHLAST						
All	2	0.26	0.52	0.00	0.00	1.03
Goal		0.52	0.73	0.52	0.00	1.03
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHSUM						
All	2	0.58	1.17	0.00	0.00	2.34
Goal		1.17	1.65	1.17	0.00	2.34
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHDUR (in seconds)						
All	2	105.88	54.06	102.88	43.00	174.75
Goal		136.25	54.45	136.25	97.75	174.75
Non-goal		75.50	45.96	75.50	43.00	108.00
RELAXED INFORMATION SCENT						
TRAILMAX						
All	5	0.10	0.33	0.00	0.00	1.04
Goal		0.21	0.47	0.00	0.00	1.04
Non-goal		0.00	0.00	0.00	0.00	0.00
TRAILLAST						
All	5	0.10	0.33	0.00	0.00	1.04
Goal		0.21	0.47	0.00	0.00	1.04
Non-goal		0.00	0.00	0.00	0.00	0.00
TRAILSUM						
All	5	0.10	0.33	0.00	0.00	1.04
Goal		0.21	0.47	0.00	0.00	1.04
Non-goal		0.00	0.00	0.00	0.00	0.00

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Table 75: Site-centric: Metric Statistics (Supported – 1.25)

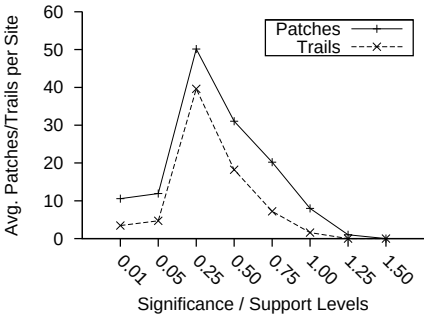
	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	1	0.65	0.92	0.65	0.00	1.31
Goal		1.31	0.00	1.31	1.31	1.31
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHLAST						
All	1	0.65	0.92	0.65	0.00	1.31
Goal		1.31	0.00	1.31	1.31	1.31
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHSUM						
All	1	0.65	0.92	0.65	0.00	1.31
Goal		1.31	0.00	1.31	1.31	1.31
Non-goal		0.00	0.00	0.00	0.00	0.00
PATCHDUR (in seconds)						
All	1	39.25	21.57	39.25	24.00	54.50
Goal		54.50	0.00	54.50	54.50	54.50
Non-goal		24.00	0.00	24.00	24.00	24.00
RELAXED INFORMATION SCENT						
TRAILMAX						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
TRAILLAST						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
TRAILSUM						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

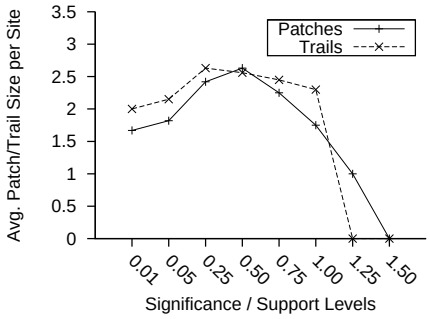
Table 76: Site-centric: Metric Statistics (Supported – 1.50)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
PATCHMAX						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
PATCHLAST						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
PATCHSUM						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
PATCHDUR (in seconds)						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
RELAXED INFORMATION SCENT						
TRAILMAX						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
TRAILLAST						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a
TRAILSUM						
All	0	n/a	n/a	n/a	n/a	n/a
Goal		n/a	n/a	n/a	n/a	n/a
Non-goal		n/a	n/a	n/a	n/a	n/a

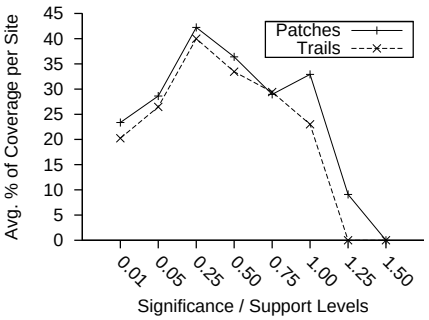
Note: all values are based on the median values from each Web site's goal and non-goal sessions.



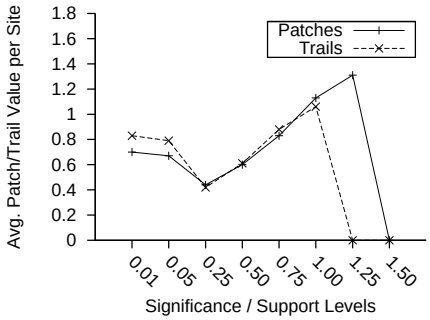
(a) Number of Patches and Trails



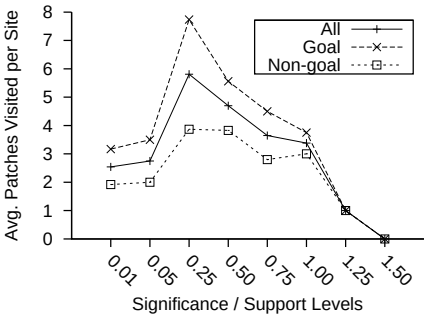
(b) Size of Patches and Trails



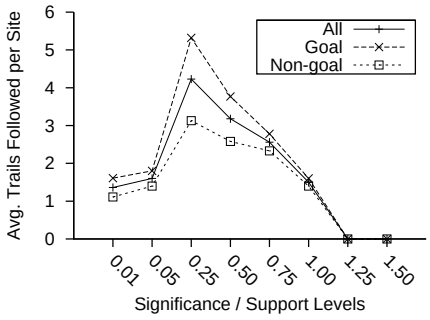
(c) % Coverage of Patches and Trails



(d) Value of Patches and Trails



(e) Patches Visited



(f) Trails Followed

Figure 61.: Site-centric: Average Patch and Trail Statistics Per Site

three support levels had patches ranging in size from 2.25 to 2.63 pages and trails from 2.45 to 2.63 pages.

The percentage of coverage for both patches and trails were nearly identical for the first five significance and support levels. The pattern of coverage roughly mirrored that of the number of patches and trails found from figure 61a. As the number of available patches and trails increased, the likelihood that more pages from a Web site may be included in a patch also increased. Thus, the increase and decrease in Web site coverage changed in a similar direction and degree as the change in discovered patches and trails. For example, the lowest number of patches and trails found along with the smallest coverage percentage was at α = 0.01 (10.58 patches with 23.37% coverage and 3.44 trails with 20.25% coverage). In contrast, the highest number of patches, trails, and coverage percentage was at support level 0.25 (50.16 patches with 42.22% coverage and 39.60 trails with 40.00% coverage).

The average value of patches and trails were relatively constant across the significance levels, but increased steadily with each support level[56]. A noticeable difference between patches and trails was only present for the two significance levels (e.g., 0.70 patch value versus 0.83 trail value at α = 0.01). In comparing the significance and support levels, there was not a direct equivalent of either significance level found within the selected support levels. For example, in order to obtain a similar value for patches and trails as found at α = 0.05 (0.67 patch value and 0.79 trail value), the support level would need to have been between 0.50 and 0.75 (0.60 – 0.83 patch value and 0.61 – 0.88 trail value). However, a support level between 0.50 and 0.75 would still not be equivalent since the value of significant patches were as low as 0.27 and 0.28 for α = 0.01 and 0.05, respectively.

The last measure (figures 61e and 61f) illustrated the number of patches visited and trails followed by foragers. Goal sessions visited more patches and followed more trails across all the different significance and support levels than non-goal sessions. In addition, the general shape of both figures followed the number of patches and trails found on a site. For example, the highest numbers of patches found and visited were both seen at support level 0.25 (50.16 patches discovered with 7.74 patches followed by goal sessions).

[56]The increase of value for each support level was not surprising since the support level created a minimum allowable value for any included patches or trails.

Table 77: Site-centric: Number of Patches by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	12	10.58	26.35	2.00	1	94
0.05	14	11.93	28.74	3.50	1	111
SUPPORTED						
0.25	32	50.16	132.79	14.50	2	748
0.50	24	31.04	83.50	6.50	1	412
0.75	10	20.20	47.69	2.00	1	155
1.00	2	8.00	8.49	8.00	2	14
1.25	1	1.00	0.00	1.00	1	1
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 78: Site-centric: Number of Trails by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	9	3.44	6.29	1.00	1	20
0.05	10	4.70	8.04	1.50	1	27
SUPPORTED						
0.25	35	39.60	97.97	12.00	1	491
0.50	24	18.21	40.69	5.00	1	188
0.75	10	7.20	11.60	3.00	1	39
1.00	5	1.60	0.89	1.00	1	3
1.25	0	n/a	n/a	n/a	n/a	n/a
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 79: Site-centric: Patch Size by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	12	1.67	0.54	1.50	1.00	3.00
0.05	14	1.82	0.64	2.00	1.00	3.00
SUPPORTED						
0.25	32	2.42	0.72	2.25	1.00	4.00
0.50	24	2.63	0.89	2.75	1.00	4.00
0.75	10	2.25	0.75	2.25	1.00	3.00
1.00	2	1.75	0.35	1.75	1.50	2.00
1.25	1	1.00	0.00	1.00	1.00	1.00
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 80: Site-centric: Trail Size by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	9	2.00	0.00	2.00	2.00	2.00
0.05	10	2.15	0.34	2.00	2.00	3.00
SUPPORTED						
0.25	35	2.63	0.65	3.00	2.00	5.00
0.50	24	2.56	0.74	2.50	2.00	5.00
0.75	10	2.45	0.50	2.25	2.00	3.00
1.00	5	2.30	0.67	2.00	2.00	3.50
1.25	0	n/a	n/a	n/a	n/a	n/a
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 81: Site-centric: Patch Coverage by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	12	23.37%	13.88%	19.62%	7.14%	50.00%
0.05	14	28.63%	13.85%	26.79%	10.00%	50.00%
SUPPORTED						
0.25	32	42.22%	16.79%	43.65%	7.89%	70.00%
0.50	24	36.39%	16.04%	35.92%	5.26%	70.00%
0.75	10	28.95%	11.66%	30.08%	12.50%	47.62%
1.00	2	32.90%	20.82%	32.90%	18.18%	47.62%
1.25	1	9.09%	0.00%	9.09%	9.09%	9.09%
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 82: Site-centric: Trail Coverage by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	9	20.25%	12.78%	18.18%	3.45%	47.62%
0.05	10	26.47%	14.80%	27.44%	6.90%	50.00%
SUPPORTED						
0.25	35	40.00%	17.18%	42.11%	2.53%	70.00%
0.50	24	33.44%	13.43%	33.33%	6.90%	55.56%
0.75	10	29.40%	14.58%	27.44%	8.33%	50.00%
1.00	5	23.02%	15.30%	18.18%	12.50%	50.00%
1.25	0	n/a	n/a	n/a	n/a	n/a
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 83: Site-centric: Patch Value by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	12	0.70	0.23	0.71	0.27	1.17
0.05	14	0.67	0.16	0.67	0.28	0.88
SUPPORTED						
0.25	32	0.44	0.11	0.43	0.28	0.67
0.50	24	0.60	0.06	0.58	0.52	0.75
0.75	10	0.83	0.12	0.78	0.75	1.17
1.00	2	1.13	0.05	1.13	1.09	1.17
1.25	1	1.31	0.00	1.31	1.31	1.31
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 84: Site-centric: Trail Value by Site

	N	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE						
0.01	9	0.83	0.21	0.86	0.50	1.06
0.05	10	0.79	0.19	0.82	0.46	1.05
SUPPORTED						
0.25	35	0.42	0.10	0.41	0.27	0.78
0.50	24	0.61	0.06	0.60	0.53	0.78
0.75	10	0.88	0.10	0.83	0.80	1.05
1.00	5	1.06	0.04	1.05	1.00	1.12
1.25	0	n/a	n/a	n/a	n/a	n/a
1.50	0	n/a	n/a	n/a	n/a	n/a

Table 85: Site-centric: Patch Visitation by Site

	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE					
0.01					
All	2.54	1.89	2.00	1.00	10.00
Goal	3.17	2.44	2.00	1.00	10.00
Non-goal	1.92	0.79	2.00	1.00	4.00
0.05					
All	2.75	2.25	2.00	1.00	11.00
Goal	3.50	2.93	2.00	1.00	11.00
Non-goal	2.00	0.88	2.00	1.00	4.00
SUPPORTED					
0.25					
All	5.81	5.88	4.00	1.00	26.00
Goal	7.74	7.58	4.00	1.00	26.00
Non-goal	3.87	2.23	4.00	1.00	9.00
0.50					
All	4.70	4.49	3.00	1.00	21.00
Goal	5.56	5.68	3.00	1.00	21.00
Non-goal	3.83	2.73	3.00	1.00	11.00
0.75					
All	3.65	3.59	2.00	1.00	14.00
Goal	4.50	4.60	2.00	1.00	14.00
Non-goal	2.80	2.10	2.00	1.00	7.00
1.00					
All	3.38	1.70	3.00	2.00	5.50
Goal	3.75	2.47	3.75	2.00	5.50
Non-goal	3.00	1.41	3.00	2.00	4.00
1.25					
All	1.00	n/a	1.00	1.00	1.00
Goal	1.00	n/a	1.00	1.00	1.00
Non-goal	1.00	n/a	1.00	1.00	1.00
1.50					
All	n/a	n/a	n/a	n/a	n/a
Goal	n/a	n/a	n/a	n/a	n/a
Non-goal	n/a	n/a	n/a	n/a	n/a

Table 86: Site-centric: Trail Following by Site

	Mean	St. Dev.	Median	Min	Max
SIGNIFICANCE					
0.01					
All	1.36	0.94	1.00	1.00	4.50
Goal	1.61	1.27	1.00	1.00	4.50
Non-goal	1.11	0.33	1.00	1.00	2.00
0.05					
All	1.60	1.14	1.00	1.00	5.00
Goal	1.80	1.48	1.00	1.00	5.00
Non-goal	1.40	0.70	1.00	1.00	3.00
SUPPORTED					
0.25					
All	4.23	5.24	3.00	1.00	34.00
Goal	5.32	6.86	3.00	1.00	34.00
Non-goal	3.13	2.49	3.00	1.00	10.00
0.50					
All	3.18	3.05	2.00	1.00	14.00
Goal	3.77	3.68	2.00	1.00	14.00
Non-goal	2.58	2.17	2.00	1.00	11.00
0.75					
All	2.56	1.62	2.50	1.00	6.00
Goal	2.78	1.86	3.00	1.00	6.00
Non-goal	2.33	1.41	2.00	1.00	4.00
1.00					
All	1.50	0.71	1.00	1.00	3.00
Goal	1.60	0.89	1.00	1.00	3.00
Non-goal	1.40	0.55	1.00	1.00	2.00
1.25					
All	n/a	n/a	n/a	n/a	n/a
Goal	n/a	n/a	n/a	n/a	n/a
Non-goal	n/a	n/a	n/a	n/a	n/a
1.50					
All	n/a	n/a	n/a	n/a	n/a
Goal	n/a	n/a	n/a	n/a	n/a
Non-goal	n/a	n/a	n/a	n/a	n/a

Hypotheses Testing

Table 87 presents a summary of the results from each of the different significance and support mining levels[57]. The table lists the hypothesis number and metric being tested in the first two columns. Columns three and four present the results when patches and trails were mined using a significance value of 0.01 and 0.05. The final six columns provide the results when the specified support level (0.25 to 1.50 in 0.25 increments) was used to learn patches and trails[58].

Table 87: Site-centric: Patches and Trails Hypotheses Results Summary

		Hypothesis Supported?							
		Significance		Support					
Hyp.	Metric	0.01	0.05	0.25	0.50	0.75	1.00	1.25	1.50
INFORMATION PATCH – PAGE-PATCH									
SC5a	PATCHMAX	Yes**	Yes***	Yes***	Yes***	Yes**	No	No	n/a
SC5b	PATCHLAST	Yes**	Yes***	Yes***	Yes***	Yes**	No	No	n/a
SC5c	PATCHSUM	Yes**	Yes***	Yes**	Yes***	Yes**	No	No	n/a
SC6	PATCHDUR	Yes***	Yes***	Yes***	Yes***	Yes**	No	No	n/a
RELAXED INFORMATION SCENT									
SC9a	TRAILMAX	Yes*	Yes**	Yes*	Yes*	Yes*	No	n/a	n/a
SC9b	TRAILLAST	Yes*	Yes**	Yes**	Yes*	Yes*	No	n/a	n/a
SC9c	TRAILSUM	Yes*	Yes**	Yes*	Yes*	Yes*	No	n/a	n/a

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

In general, the results for all seven measures appeared to hold fairly steady across the first five significance and support mining levels. The PATCHMAX and PATCHLAST measures both had the same pattern of significant findings. The metrics were significant at α = 0.01 for all but the most stringent significance (0.01) and support mining levels (0.75), where the measures were both significant at α = 0.05. PATCHSUM followed a similar pattern as the other two patch visitation measures. However, unlike PATCHMAX and PATCHLAST, PATCHSUM was only significant at α = 0.05 for patches mined at the 0.25 support level. The drop in significance may be a symptom of the

[57]The results were determined from the sign test. As the data used for the sensitivity analysis was from the same data set that was used to test the site-centric model, the assumptions of the sign test still held.

[58]The analysis of results does not include supported levels greater than 0.75. There were too few Web sites at those mined support levels to possibly obtain statistically significant results.

patches found at the 0.25 support mining level covering too much of a Web site (42.22% average coverage) to be as effective at distinguishing between goal and non-goal sessions.

The PATCHDUR metric was significant at α = 0.01 for all levels except the 0.75 support mining level, where the measure was significant at α = 0.05. The decrease in significance may be due to the sign test's lack of power in detecting differences at α = 0.01 with a sample size of only 10 Web sites.

TRAILMAX, TRAILLAST, and TRAILSUM were all significant at α = 0.10 except for trails mined at the 0.05 significance level, where all three measures were significant at α = 0.05. In addition, the TRAILLAST measure was also significant at α = 0.05 at the 0.25 support mining level. A lack of power by the sign test to adequately detect a difference in such small sample sizes (e.g., five to nine Web sites) was the primary suspect for many of the measures only reaching a significance of α = 0.10.

Tables 88 – 95 present the results of all eight significance and support mining levels for all three statistical tests. Following the tables, figure 62 illustrates the p-values obtained from the statistical tests for each of the seven measures. The graphs show the results of the three tests over the first five significance and support mining levels (0.01 – 0.75).

Table 88: Site-centric: Results (Significant – 0.01)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	12	8	3.72	11	0.0017^{***}	36	0.0039^{**}	8	0.0039^{**}
SC5b	PATCHLAST	12	8	3.90	11	0.0012^{***}	36	0.0039^{**}	8	0.0039^{**}
SC5c	PATCHSUM	12	8	2.92	11	0.0070^{**}	36	0.0039^{**}	8	0.0039^{**}
SC6	PATCHDUR	12	12	3.93	11	0.0012^{***}	78	0.0002^{***}	12	0.0002^{***}
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	9	5	2.92	8	0.0096^{**}	15	0.0313^{*}	5	0.0313^{*}
SC9b	TRAILLAST	9	5	2.94	8	0.0094^{**}	15	0.0313^{*}	5	0.0313^{*}
SC9c	TRAILSUM	9	5	2.57	8	0.0165^{**}	15	0.0313^{*}	5	0.0313^{*}

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 89: Site-centric: Results (Significant – 0.05)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	14	9	3.68	13	0.0014^{***}	45	0.0020^{***}	9	0.0020^{***}
SC5b	PATCHLAST	14	9	3.92	13	0.0009^{***}	45	0.0020^{***}	9	0.0020^{***}
SC5c	PATCHSUM	14	9	3.00	13	0.0051^{**}	45	0.0020^{***}	9	0.0020^{***}
SC6	PATCHDUR	14	14	4.11	13	0.0006^{***}	100	0.0006^{***}	13	0.0009^{***}
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	10	6	3.33	9	0.0044^{**}	21	0.0156^{**}	6	0.0156^{**}
SC9b	TRAILLAST	10	6	3.36	9	0.0042^{**}	21	0.0156^{**}	6	0.0156^{**}
SC9c	TRAILSUM	10	6	2.89	9	0.0089^{**}	21	0.0156^{**}	6	0.0156^{**}

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 90: Site-centric: Results (Supported – 0.25)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	32	20	3.49	31	0.0007^{***}	191	0.0003^{***}	17	0.0013^{***}
SC5b	PATCHLAST	32	17	3.36	31	0.0011^{***}	141	0.0005^{***}	15	0.0012^{***}
SC5c	PATCHSUM	32	21	3.04	31	0.0024^{***}	210	0.0002^{***}	17	0.0036^{**}
SC6	PATCHDUR[a]	31	31	4.20	30	0.0001^{***}	450	$< 0.0001^{***}$	27	$< 0.0001^{***}$
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	35	20	2.49	34	0.0089^{**}	167	0.0096^{**}	15	0.0207^{*}
SC9b	TRAILLAST	35	19	2.23	34	0.0162^{**}	153	0.0090^{**}	15	0.0096^{**}
SC9c	TRAILSUM	35	20	2.33	34	0.0128^{**}	181	0.0016^{***}	15	0.0207^{*}

[a] PATCHDUR only had a total of 31 Web sites (versus the 32 sites in PATCHES) because there were not any sessions which visited discovered goal patches at one Web site. All five discovered goal patches at the site of interest contained a page that was no longer available to sessions within the testing set. More specifically, the training set consisted of sessions which existed on or before 05/23/2008 8:29:38 PM. The page in question was last visited by any session on 03/20/2008 8:13:05 PM.

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 91: Site-centric: Results (Supported – 0.50)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	24	9	3.40	23	0.0012^{***}	45	0.0020^{***}	9	0.0020^{***}
SC5b	PATCHLAST	24	9	3.36	23	0.0014^{***}	45	0.0020^{***}	9	0.0020^{***}
SC5c	PATCHSUM	24	9	2.84	23	0.0047^{**}	45	0.0020^{***}	9	0.0020^{***}
SC6	PATCHDUR	24	23	2.80	23	0.0050^{***}	227	0.0027^{***}	18	0.0053^{***}
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	24	9	2.45	23	0.0112^{**}	41	0.0137^{**}	8	0.0195^{*}
SC9b	TRAILLAST	24	9	2.12	23	0.0226^{*}	39	0.0273^{*}	8	0.0195^{*}
SC9c	TRAILSUM	24	9	2.29	23	0.0159^{**}	42	0.0098^{**}	8	0.0195^{*}

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 92: Site-centric: Results (Supported – 0.75)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	10	6	3.50	9	0.0034^{**}	21	0.0156^{**}	6	0.0156^{**}
SC5b	PATCHLAST	10	6	3.61	9	0.0028^{***}	21	0.0156^{**}	6	0.0156^{**}
SC5c	PATCHSUM	10	6	2.70	9	0.0123^{**}	21	0.0156^{**}	6	0.0156^{**}
SC6	PATCHDUR	10	10	2.62	9	0.0138^{**}	50	0.0098^{***}	9	0.0107^{**}
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	10	5	2.83	9	0.0099^{**}	15	0.0313^{*}	5	0.0313^{*}
SC9b	TRAILLAST	10	5	2.80	9	0.0104^{**}	15	0.0313^{*}	5	0.0313^{*}
SC9c	TRAILSUM	10	5	2.58	9	0.0148^{**}	15	0.0313^{*}	5	0.0313^{*}

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 93: Site-centric: Results (Supported – 1.00)

		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	2	1	1.00	1	0.2500	1	0.5000	1	0.5000
SC5b	PATCHLAST	2	1	1.00	1	0.2500	1	0.5000	1	0.5000
SC5c	PATCHSUM	2	1	1.00	1	0.2500	1	0.5000	1	0.5000
SC6	PATCHDUR	2	2	10.13	1	0.0313**	3	0.2500	2	0.2500
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	5	1	1.00	4	0.1870	1	0.5000	1	0.5000
SC9b	TRAILLAST	5	1	1.00	4	0.1870	1	0.5000	1	0.5000
SC9c	TRAILSUM	5	1	1.00	4	0.1870	1	0.5000	1	0.5000

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 94: Site-centric: Results (Supported – 1.25)

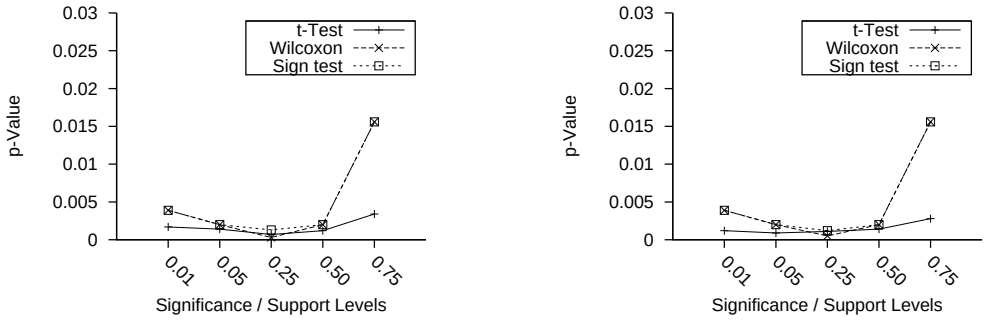
		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	1	1	n/a	n/a	n/a	1	0.5000	1	0.5000
SC5b	PATCHLAST	1	1	n/a	n/a	n/a	1	0.5000	1	0.5000
SC5c	PATCHSUM	1	1	n/a	n/a	n/a	1	0.5000	1	0.5000
SC6	PATCHDUR	1	1	n/a	n/a	n/a	1	0.5000	1	0.5000
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC9b	TRAILLAST	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC9c	TRAILSUM	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

Table 95: Site-centric: Results (Supported – 1.50)

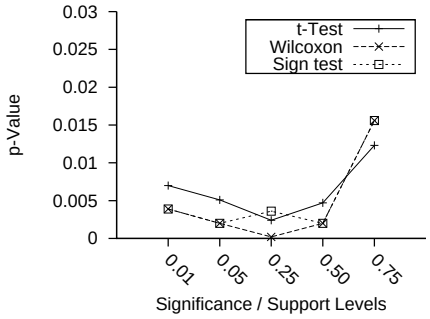
		N		t-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Zeros	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
SC5a	PATCHMAX	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC5b	PATCHLAST	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC5c	PATCHSUM	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC6	PATCHDUR	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
RELAXED INFORMATION SCENT										
SC9a	TRAILMAX	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC9b	TRAILLAST	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
SC9c	TRAILSUM	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

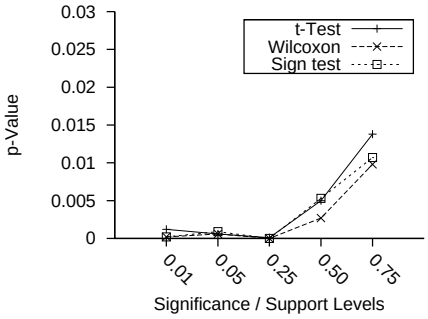


(a) PATCHMAX

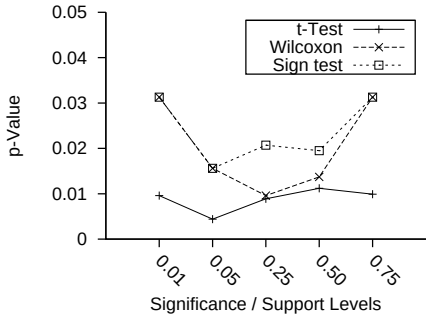
(b) PATCHLAST



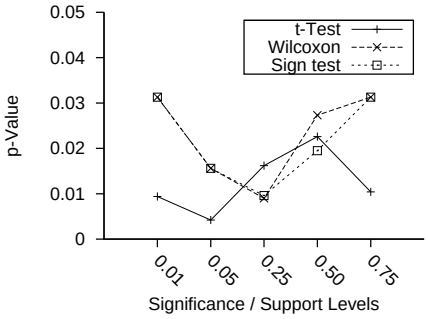
(c) PATCHSUM



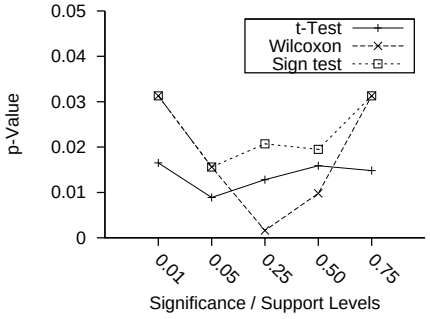
(d) PATCHDUR



(e) TRAILMAX



(f) TRAILLAST



(g) TRAILSUM

Figure 62.: Site-centric: Trail and Patch p-values by Significance / Support Levels

7.3 Conclusion

This chapter provided the results of both the user- and site-centric models of information foraging. Descriptive statistics, assumption checks of statistical tests, and results from each model's hypotheses were provided. In addition, a sensitivity analysis was done on the seven hypotheses of the site-centric model that relied on mining patches and trails. Overall, three of the four user-centric hypotheses were supported at α = 0.01. Of the 13 site-centric hypotheses and sub-hypotheses, seven were supported at α = 0.01, four at α = 0.05, one at α = 0.10, and one at α = 0.01 in the opposite direction as expected.

Chapter 8

Temporal Aspects of Information Foraging

The site-centric clickstream model of information foraging made an implicit assumption that the structure of the Web sites being examined did not change over the course of the analysis. Thus, it was expected that browsing patterns of goal and non-goal sessions would be roughly constant over time, for both the calculated measures (e.g., duration, number of pages viewed) and learned patches and trails. However, the Web is a dynamic and evolving environment (Warren et al., 1999; Chi et al., 1998) where Web sites add, modify, and remove content (Pitkow and Pirolli, 1997) on a regular basis[1]. In addition, like traditional software, Web sites may also undergo structural maintenance to improve the quality of the browsing experience for visitors (Ricca and Tonella, 2001).

As Web sites can be dynamic, assuming a static representation may not be appropriate when testing the site-centric model[2]. Therefore, this chapter presents a second test of the site-centric hypotheses using temporal aspects to determine if time makes a difference in the results. Instead of comparing browsing behavior to an absolute point of zero, all behavior was compared relative to prior goal sessions at the site of interest. Thus, if content or structural changes occurred, they would be reflected in the relative value of the current session.

Methodologically, relative measures were determined by progressively calculating sessions in order of their session start time. Thus, the currently processed session would be compared relative to all goal sessions that occurred before it. Although the comparison was relatively simple in definition (i.e., all prior goal sessions were used as opposed to a sliding window), the computational complexity of the methodology was still much higher than for the static site-centric version. Therefore, the results of this chapter may also shed light onto the value of undertaking the extra complexity of this methodology.

[1] For example, Ricca and Tonella (2000) analyzed 15 Web sites over a three month period and found that each Web site had, on average, 3.4 significant structural changes within that time frame.

[2] The same concern over static Web sites was not an issue in the tested portions of the user-centric model. Comparisons were made relative to browsing behavior at other Web sites within the limited time of the user's session. Thus, the only expectation was that the Web site would remain static while the user was on the site of interest.

The first section of this chapter details the methodology used to test the temporal version of the site-centric model. In particular, the data used to test the model is explained first, followed by the algorithm used to progressively calculate measures, and then finally the formulas used to determine the relative value for each session's measures. The second section presents the results of the temporal version and compares against the results of the static version of the model. Finally, the conclusion summarizes the usefulness of the temporal methodology given the results obtained.

8.1 Methodology

In the first subsection below, a description of the data elements available in the data are described[3]. The second subsection details the progressive manner in which the dataset was processed. The processing was done in order to create measures that were relative to prior sessions. Finally, the last subsection illustrates the equations used to calculate each of the measures for the temporal version of the site-centric model.

8.1.1 Dataset Sample

The data used in the static and temporal versions of site-centric model were exactly the same. The data contained a set of n sessions S (S_0, S_1, ..., S_{n-1}), where S_i represents a single session. Each session (S_i) contained a set of m page information tuples P (P_{i0}, P_{i1}, ..., P_{im-1}), where P_{ij} represents information about a particular page viewed during a session. Each page information tuple was made up of seven pieces of information: a unique identifier for the session, Web site, referring domain, and page viewed; date and time the page was viewed; how much time was spent on the page; and if the page represented a contact goal being achieved.

The calculation of metrics for each session was only done on those parts of a session occurring *before* the achievement of a contact goal[4]. This truncation was done because the problem being investigated was the prediction of goal achievement during the *remainder* of a session. Thus, prediction was done from a point right before a form submission occurred, i.e., P *only* contained pages which occurred *before* the contact form was submitted for the contact goal of interest.

[3] Summary statistics about the site-centric dataset can be found in chapter 6.
[4] If a session did not submit a contact form then the entire session was used.

8.1.2 Progressive Calculations

The measures in the temporal version of the site-centric model compared the browsing behavior of the current forager against what previous goal-achieving foragers had done[5]. For calculating the measures, previous was defined as any session that started before the current forager's session began. For example, a session which took place a month after data collection began would have had that entire month's worth of goal sessions to compare against. For a session that took place six months after the start of data collection, there would have been even more goal sessions to compare against.

The process used to compare prior sessions against is outlined in the *processDataset* algorithm (figure 63). The algorithm requires two arguments: a set of sessions for a particular Web site and the minimum percentage of goal sessions to bank before the calculation of sessions' measures should begin. The *processDataset* algorithm operates in six basic steps.

The first step of the *processDataset* algorithm sorted sessions in ascending order by their start date and time (line 23). The second step (line 25) determined the total number of goal sessions in the entire set of sessions. After setting up the environment in the first two steps, each session was then iterated over (lines 27-45) for the next three steps.

The third step (line 29) determined if the minimum percentage of goal sessions had been added to the set of banked goal sessions. If the minimum percentage had been met, then the measures for the current session were calculated and added to the dataset (line 30). Calculations were performed using all banked goal sessions along with valuable goal patches and trails. The fourth step then added the current session into the appropriate set of banked sessions in lines 34-38. A session was banked regardless of if measures were calculated for that session or not.

The sixth step handled the mining of goal patches and trails (lines 41-44). Since long tail sites have limited data and an additional goal session may have an impact on the formation of patches or trails, mining was done after each goal session was added to the bank (once the minimum percentage was met)[6]. The final step occurred after all sessions had been processed. In the last step,

[5]Previous non-goal sessions were also used, but only for learning patches and trails. See §8.1.3 for more details.

[6]Patches and trails were not mined after every non-goal session because there were generally many more non-goal sessions than goal sessions. Thus, the addition of one additional non-goal session, when finding frequent itemsets or sequential patterns, was unlikely to cause drastic differences in patch and trail formation, unlike what may occur with goal sessions. In addition, the computational effort required to mine after every session would be very high at some Web sites (e.g., a site with 40,000 sessions).

```
1  /**
2   * Parameters: (a) Set of user sessions: S = {S_0, S_1, ..., S_{N-1}}
3   *                 where S_i = set of page information tuples, P.
4   *             (b) Minimum percentage of goal sessions
5   *                 to base calculations on: minPercent
6   * Returns:    Set of result records: R = {R_0, R_1, ..., R_{X-1}}
7   * Methods:    (a) calculateMeasures(S_i, G, A, T): calculates
8   *                 all the necessary measures for session S_i using
9   *                 all goal sessions from set G, goal patches from set A,
10  *                 and goal trails from set T
11  *             (b) generatePatches(G, N): returns a set of valuable goal patches
12  *             (c) generateTrails(G, N): returns a set of valuable goal trails
13  *             (d) getGoalCount(S): returns number of goal sessions in S
14  *             (e) isGoal(S_i): true if session achieved goal
15  *             (f) sort(S): sorts sessions in ascending order by session
16  *                 start date
17  */
18 processDataset(S, minPercent) {
19   // R = result records; G = banked goal sessions; N = banked non-goal sessions
20   // A = valuable goal patches; T = valuable goal trails
21   R = {}; G = {}; N = {}; A = {}; T = {};
22
23   sort(S); // sort sessions in ascending order by session start date
24
25   goalCount = getGoalCount(S); // determine how many total goal sessions in entire set
26
27   for each (i ∈ S) {
28     // Only calculate once minimum percentage of goal sessions is met
29     if (||G|| / goalCount >= minPercent) {
30       R += calculateMeasures(i, G, A, T);
31     }
32
33     // Add session to banked goal or non-goal set
34     if (isGoal(i)) {
35       G += i;
36     } else {
37       N += i;
38     }
39
40     // Mine patches and trails for each new goal session (if enough goal sessions are banked)
41     if (isGoal(i) && ||G|| / goalCount >= minPercent) {
42       A = generatePatches(G, N);
43       T = generateTrails(G, N);
44     }
45   }
46   return R;
47 }
```

Figure 63.: Temporal Site-centric: processDataset Algorithm

the set of result data records which contained the calculated measures for each session were returned. Of note is the algorithm did not include those sessions in the returned results that occurred before the minimum percentage of sessions was met.

Example

Table 96 presents an example of how the algorithm processed a dataset. The table shows the first 11 sessions from the dataset sorted by session start time. Five of the sessions resulted in a goal being achieved. All of the sessions were passed to the algorithm. In addition, the minimum percentage of goal sessions required before calculating measures was set to 80%. Therefore, sessions were not considered part of the result dataset until four goal sessions (80%) were banked.

Table 96: Temporal Site-centric: Example Sessions

Session	Start Date and Time	Goal Achieved?
S1	7/09/08 11:04:07	No
S2	7/09/08 17:35:12	Yes
S3	7/11/08 10:10:56	No
S4	7/15/08 11:36:18	Yes
S5	7/15/08 11:37:08	Yes
S6	7/15/08 14:43:23	No
S7	7/22/08 12:11:10	No
S8	7/23/08 19:44:39	Yes
S9	7/23/08 20:23:21	No
S10	7/25/08 14:05:09	Yes
S11	7/26/08 16:07:25	No
	⋮	

Table 97 illustrates the process using the sessions from table 96. The contents of which sessions were in the result set, goal set, and non-goal set are provided at the *end* of every iteration of the algorithm (i.e., line 45). Calculations for a session were done *before* the session was added to either the goal or non-goal set.

After processing the first session the result and goal set remained empty while $S1$ was added to

the non-goal set. After the eighth session was processed the minimum percentage of goal sessions was met for the goal set. Sessions $S2$, $S4$, $S5$, and $S8$ were included in the goal set while sessions $S1$, $S3$, $S6$, and $S7$ were in the non-goal set. Up to the eighth session, no sessions had been added to the result data set yet (i.e., no calculations had been performed).

Table 97: Temporal Site-centric: Example Dataset Processing

Step	Result Set	Goal Set	Non-Goal Set
1			$S1$
2		$S2$	$S1$
3		$S2$	$S1, S3$
4		$S2, S4$	$S1, S3$
5		$S2, S4, S5$	$S1, S3$
6		$S2, S4, S5$	$S1, S3, S6$
7		$S2, S4, S5$	$S1, S3, S6, S7$
8		$S2, S4, S5, S8$	$S1, S3, S6, S7$
9	$S9$	$S2, S4, S5, S8$	$S1, S3, S6, S7, S9$
10	$S9, S10$	$S2, S4, S5, S8, S10$	$S1, S3, S6, S7, S9$
11	$S9, S10, S11$	$S2, S4, S5, S8, S10$	$S1, S3, S6, S7, S9, S11$
		$\vdots$	

After the eighth step; however, the minimum percentage of goal sessions had been met. Therefore, all remaining sessions would have their measures calculated and added to the result data set. Session $S9$ used the patches and trails mined from the four banked goal ($S2, S4, S5, S8$) and non-goal sessions ($S1, S3, S6, S7$), along with just the banked goal sessions to calculate its relative measures. For session $S10$, the previous session ($S9$) was added to the non-goal set, but the patches and trails were not re-mined. For the final session, new patches and trails were mined, because $S10$ was a goal session. If more than eleven sessions existed, then this progressive manner of mining patches and trails and calculating measures would have continued until the final session was processed.

In this research the *processDataset* algorithm was run with the minimum percentage of goal sessions set to 70%. Thus, measures were only calculated when at least 70% of all *goal* sessions were banked.

8.1.3 Metrics

Table 98 summarizes the metrics used to test the temporally-positioned hypotheses for the site-centric clickstream model (TSC). The name of each metric along with a description of how it was calculated is provided. In addition, the hypothesis which corresponds to the metric is also provided in the table. A more in-depth description of the metrics is given in the following subsections.

Table 98 does not contain the RETURN and VISITED metrics (hypotheses SC3 and SC4) because they were calculated at a Web site as opposed to an individual level of analysis. The temporal version of the model examines relative behavior of a *user* versus previous sessions. Therefore, measures at a higher level of analysis were not analyzed.

To help clarify the notation being used below for the metrics, C represents the current session being analyzed, G is the set of banked past goal sessions that C will be compared against, and $median()$ is a function that returns the median from a set of values.

Information Patch – Site-Patch

RELDUR is the total duration in seconds a visitor has spent at a Web site relative to the median time prior goal sessions have spent at the same Web site. The relative duration is calculated from equation 8.1, where $duration(i)$ is the duration spent during session i. To obtain RELDUR, the median duration of all banked goal sessions in the goal set G is subtracted from the total duration of the current session C.

$$\text{RELDUR} = duration(C) - median\left(\textbf{for each}_{i \in G}\left[duration(i)\right]\right) \tag{8.1}$$

RELPGS is the number of pages a visitor has viewed at a Web site relative to the median number of pages viewed by prior goal sessions at the same Web site. The relative number of pages is calculated as shown in equation 8.2, where $pages(i)$ is the number of pages viewed during session i. To acquire RELPGS, the median number of pages viewed from all goal sessions in goal set G is subtracted from the number of pages viewed during the current session C.

$$\text{RELPGS} = pages(C) - median\left(\textbf{for each}_{i \in G}\left[pages(i)\right]\right) \tag{8.2}$$

Table 98: Temporal Site-centric: Model Metrics

Hypothesis #	Metric	Description
INFORMATION PATCH – SITE-PATCH		
TSC1	RELDUR	Duration in seconds spent on a Web site relative to past goal sessions.
TSC2	RELPGS	Number of pages viewed on a Web site relative to past goal sessions.
INFORMATION PATCH – PAGE-PATCH		
TSC5a	RELPTCMAX	Maximum value of any goal page-patch visited relative to past goal sessions.
TSC5b	RELPTCLAST	Value of last goal page-patch visited relative to past goal sessions.
TSC5c	RELPTCSUM	Total value of all goal page-patches visited relative to past goal sessions.
TSC6	RELPTCDUR	Median duration in seconds spent in all goal page-patches relative to past goal sessions.
STRICT INFORMATION SCENT		
TSC7	RELUNQ	Percentage of unique pages viewed relative to past goal sessions.
TSC8	RELLNR	Linearity of clickstream relative to past goal sessions.
RELAXED INFORMATION SCENT		
TSC9a	RELTRLMAX	Maximum value of any goal trail followed relative to past goal sessions.
TSC9b	RELTRLLAST	Value of last goal trail followed relative to past goal sessions.
TSC9c	RELTRLSUM	Total value of all goal trails followed relative to past goal sessions.
OTHER		
n/a	GOAL	Whether a goal occurred during the session.

Information Patch – Page-Patch

Patches at a Web site must already be known in order to calculate the four RELPTC visitation metrics: RELPTCMAX, RELPTCLAST, RELPTCSUM, and RELPTCDUR. The methodology for learning patches is described in detail in appendix 5.B. In general, learning patches requires a set of goal and non-goal sessions to determine which parts of a Web site (i.e., pages) are better able to distinguish between the two groups. Patches are specific to a single Web site.

As the four RELPTC metrics require patches to be learned first in order to quantify a session's patch visitation, the banked goal and non-goal sessions (G and N) were used to discover goal patches at a Web site. The current session then calculated the RELPTC metrics from the learned goal patches. However, the current session would only calculate the RELPTC metrics *if and only if* goal patches were found at the Web site. In addition, the RELPTCDUR metric would only be calculated for the current session *if and only if* that session visited at least one of the goal patches discovered at the Web site of interest. Furthermore, since the measures for the temporal site-centric model are all relative to prior goal sessions, the same goal sessions used to learn the patches also calculated the RELPTC metrics for their own respective sessions so that relative comparisons could be made.

Learning Patches

Patches were learned for a Web site using the training dataset (R), which consisted of banked goal (G) and non-goal (N) sessions, according to the methodology outlined in appendix 5.B. Patches were learned at an α level of 0.05[7].

Specifically, a set of n valuable patches A (A_0, A_1, …, A_{n-1}) were discovered, where A_i represents a single valuable patch. A_i consists of a set of m unordered and distinct pages U (U_0, U_1, …, U_{m-1}).

Each patch (A_i) was also given a value according to equation 8.3 (Yang and Padmanabhan, 2003). S_{Gi} and S_{Ni} represent the number of goal and non-goal sessions from the training dataset that visited patch A_i, respectively. R_G and R_N is the total number of goal and non-goal sessions from the training dataset. The value of patch A_i could range from zero to two, with higher num-

[7] A more in-depth description of learning patches may be found in §5.2.2.

bers representing a greater difference in support of the patch in distinguishing between goal and non-goal sessions (i.e., being more valuable).

$$value(A_i) = \frac{\left| \frac{S_{Gi}}{R_G} - \frac{S_{Ni}}{R_N} \right|}{\frac{1}{2}\left(\frac{S_{Gi}}{R_G} + \frac{S_{Ni}}{R_N} \right)} \tag{8.3}$$

Calculating RELPTC Metrics

To calculate the RELPTC metrics for a given session, two steps were required. First, it was determined what patches the session visited from the set of valuable patches (A). Each session had a set of l visited patches V (V_0, V_1, ..., V_{l-1}), where V_j was an individual patch visited by the current session. A session was considered to have visited a patch if all pages of the patch (U) were visited at least once (in any order) by the current session (as determined by the set of pages P from the session). Formally, A_i was added to V if $U \subseteq P$. Once it was known what patches were visited, then the four measures were calculated.

PATCHMAX is the value of the most valuable patch visited by the current user. The maximum value is determined by iterating over every visited patch to find the one with the highest value (equation 8.4). If the user did not visit any patches then the value of PATCHMAX would be zero.

$$\text{PATCHMAX} = \begin{cases} max\left(\textbf{for each}_{j \in V}(value(V_j))\right) & \text{if } \|V\| > 0 \\ 0 & \text{else} \end{cases} \tag{8.4}$$

RELPTCMAX was calculated as shown in equation 8.5. The median PATCHMAX value of all banked goal sessions in the goal set G was subtracted from the current session's (C) value of PATCHMAX in order to calculate RELPTCMAX.

$$\text{RELPTCMAX} = \text{PATCHMAX}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{PATCHMAX}(i)\right]\right) \tag{8.5}$$

PATCHLAST is the value of the last patch visited by the user[8]. Equation 8.6 illustrates how RELPTCLAST was calculated. The median PATCHLAST value of all banked goal sessions from the goal set G is subtracted from the current session's (C) value of PATCHLAST to arrive at RELPTCLAST.

[8]Details on the four-step heuristic used to determine which patch was visited last during a user's sessions may be found in §5.2.2.

$$\text{RELPTCLAST} = \text{PATCHLAST}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{PATCHLAST}(i)\right]\right) \tag{8.6}$$

PATCHSUM adds up the value of every patch visited by the current user (equation 8.7). A value of zero is given to any user that did not visit any patches.

$$\text{PATCHSUM} = \begin{cases} \sum_{j \in V}\left(value(V_j)\right) & \text{if } \|V\| > 0 \\ 0 & \text{else} \end{cases} \tag{8.7}$$

Equation 8.8 illustrates how RELPTCSUM was calculated. The metric was determined by subtracting PATCHSUM for the current session C from the median PATCHSUM value of all banked goal sessions in the goal set G.

$$\text{RELPTCSUM} = \text{PATCHSUM}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{PATCHSUM}(i)\right]\right) \tag{8.8}$$

PATCHDUR is the median duration a user spent in all their visited patches. Only sessions which visited at least one patch (i.e., $\|V\| > 0$) would have a value for PATCHDUR. The calculation for PATCHDUR is shown in equation 8.9. $totalTime(k, P)$ returns the total time a session with pages P spent on page k. If a session visited page k more than once in P, then the sum duration from all k page visitations was returned.

$$\text{PATCHDUR} = median\left[\textbf{for each}_{j \in V}\left(\sum_{k \in G} totalTime(k, P)\right)\right] \tag{8.9}$$

The manner in which RELPTCDUR was calculated is shown in equation 8.10. To obtain RELPTCDUR, the median PATCHDUR value of all banked goal sessions in the goal set G was subtracted from the current session's (C) value of PATCHDUR.

$$\text{RELPTCDUR} = \text{PATCHDUR}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{PATCHDUR}(i)\right]\right) \tag{8.10}$$

Strict Information Scent

UNIQUE is the percentage of unique pages viewed during a session. The percentage of unique pages viewed for the current visitor is calculated according to equation 8.11, where $distinct(P)$ is

the number of distinct pages viewed in the set of page information tuples P.

$$\text{UNIQUE} = \left(\frac{distinct(P)}{\|P\|} \right) * 100 \tag{8.11}$$

The relative percentage of unique pages RELUNQ is determined by subtracting the median UNIQUE value of all banked goal sessions (G) from the value of the current session's UNIQUE (equation 8.12).

$$\text{RELUNQ} = \text{UNIQUE}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{UNIQUE}(i)\right]\right) \tag{8.12}$$

LINEAR is the complexity of a session as calculated via the stratum measure. Complexity is determined via the straightness (i.e., absence of visiting pages repeatedly) of a user's browsing behavior, where higher linearity equates to less complexity. Stratum is a measure of linearity from graph theory (McEneaney, 2001) and details on its calculation may be found in appendix 5.A. RELLNR was calculated according to equation 8.13, where the median LINEAR value from the banked goal set (G) was subtracted from the current session's value of LINEAR.

$$\text{RELLNR} = \text{LINEAR}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{RELLNR}(i)\right]\right) \tag{8.13}$$

Relaxed Information Scent

The three RELTRL metrics for the relaxed information scent were calculated in a very similar manner as the RELPTC metrics. The same training set used to discover patches was used to learn trails. Both the current session and the goal sessions from the training set then used those learned trails to calculate their values for the three RELTRL metrics.

Specifically, a set of n valuable trails T (T_0, T_1, …, T_{n-1}) were discovered from the training set, where T_i represents a single valuable trail. T_i consists of a set of m *ordered* pages O (O_0, O_1, …, O_{m-1}), where the pages may repeat themselves in the ordered set (e.g., $\langle A, B, B, A, C \rangle$). Once discovered, trails were given a value like patches using equation 8.3 (with T_i being used instead of A_i).

Once the trails were discovered, each session required two steps to calculate the RELTRL measures. First, it was determined what trails were followed by the session of interest from the set of

valuable trails (T). Each session had a set of l followed trails F (F_0, F_1, ..., F_{l-1}), where F_j was an individual trail followed by the current session. A session was considered to have followed a trail if all pages of the trail (O) were followed in order by the current session (as determined by the set of pages P from the session). Although all pages must have been followed in order, repeat visitation and gaps between pages were allowed (i.e., other pages may be visited in between pages from the trail). More specifically, T_i was added to F if $O \subseteq P$ and the pages of O were found in the same order in P. Once it was known what trails were followed, then the three measures were calculated.

TRAILMAX is the value of the most valuable followed trail by the current user. The maximum value is determined by iterating over every followed trail to find the one with the highest value (equation 8.14). If the user did not visit any trails then the value of TRAILMAX would be zero.

$$\text{TRAILMAX} = \begin{cases} max\left(\textbf{for each}_{j \in F}(value(F_j))\right) & \text{if } \|F\| > 0 \\ 0 & \text{else} \end{cases} \tag{8.14}$$

RELTRLMAX was calculated as shown in equation 8.15, where the median TRAILMAX value of all banked goal sessions in the goal set G was subtracted from the current session's (C) value of TRAILMAX.

$$\text{RELTRLMAX} = \text{TRAILMAX}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{TRAILMAX}(i)\right]\right) \tag{8.15}$$

TRAILLAST is the value of the last trail followed by the user[9]. Equation 8.16 illustrates how RELTRLLAST was calculated. The median TRAILLAST value of all banked goal sessions from the goal set G is subtracted from the current session's (C) value of TRAILLAST to arrive at RELTRLLAST.

$$\text{RELTRLLAST} = \text{TRAILLAST}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{TRAILLAST}(i)\right]\right) \tag{8.16}$$

TRAILSUM adds up the value of every followed trail by the current user (equation 8.17). A value of zero is given to any user that did not visit any trails.

[9]Details on the four-step heuristic used to determine which trail was followed last during a user's sessions may be found in §5.2.2.

$$\text{TRAILSUM} = \begin{cases} \sum_{j \in F} (value(F_j)) & \text{if } \|F\| > 0 \\ 0 & \text{else} \end{cases} \tag{8.17}$$

Equation 8.18 illustrates how RELTRLSUM was calculated. The metric was determined by subtracting TRAILSUM for the current session C from the median TRAILSUM value of all banked goal sessions in the goal set G.

$$\text{RELTRLSUM} = \text{TRAILSUM}(C) - median\left(\textbf{for each}_{i \in G}\left[\text{TRAILSUM}(i)\right]\right) \tag{8.18}$$

Other

The mutually exclusive binomially distributed metric GOAL specifies whether at some point during the remainder of a session a contact form was submitted for the contact goal of interest. If a goal will be achieved during the session, GOAL will have the value of *true*. Otherwise, GOAL will have a value of *false*.

8.2 Results

The temporal site-centric model consisted of seven hypotheses about information scent and trails. Descriptive statistics of the dataset and each measure are provided in the first subsection below. The results for each of the seven hypotheses are then provided in the next subsection.

8.2.1 Descriptive Statistics

Table 99 presents the mean, standard deviation, median, minimum, and maximum number of sessions per Web site in three categories: all, goal, and non-goal sessions. Statistics for the entire dataset are shown first, followed by the number of sessions initially used in the training set. The training set first contained all sessions occurring before the first 70% of goal sessions. However, since measures were calculated in a progressive manner, the training set increased in size after each processed session.

The training set (or set of banked sessions), was used to calculate the measures for each session after the minimum percent of goal sessions was reached. A total of 3,744.24 sessions (70.35%)

Table 99: Temporal Site-centric: Sessions by Site

	Mean	St. Dev.	Median	Min	Max
ENTIRE DATASET					
All	5,322.60	7,473.76	2,637.00	245	44,405
Goal	105.94	90.13	79.00	51	587
Non-goal	5,216.66	7,427.53	2,566.00	192	44,111
MINIMUM TRAINING SET					
All	3,744.23	5,418.42	1,696.00	168	31,730
Goal	74.28	63.07	56.00	36	411
Non-goal	3,669.96	5,386.00	1,656.00	130	31,525

per Web site, on average, had their measures calculated in a progressive manner from prior goal sessions. New patches and trails were learned on each Web site over thirty different times (30.66). Each addition mining procedure also meant that all previous goal sessions had to recalculate their RELPTC and RELTRL measures against the new patches and trails.

Table 100 displays the mean, standard deviation, median, minimum, and maximum values for each of the four measures that did not require mining of patches and trails. The statistics are broken down into three groups of sessions: all, goal, and non-goal. The same 47 Web sites used in the site-centric version were also used in the temporal version.

The average relative duration of all users was 2.10 fewer minutes on a site than previous goal sessions. Goal sessions spent 0.27 more minutes than past goal sessions on a site, while non-goal sessions spent 4.46 fewer minutes. A pattern similar to the relative duration of time between the three groups was also seen for the relative number of pages. Amongst all foragers, 0.15 fewer pages were viewed on average compared to prior goal sessions. Goal sessions viewed relatively more pages than non-goal sessions did (0.12 versus -0.41) when compared to prior goal sessions.

All three groups viewed a lower percentage of unique pages, on average, than past goal sessions: -11.67% for all, -2.33% for goal, and -21.02% for non-goal. Although the average was negative for goal sessions, the median value shows goal sessions had exactly the same percentage of unique pages viewed as past sessions (i.e., 0.00%)[10].

[10]The negative relative value for percentage of unique pages may have also been a symptom of the evolution of Web sites. For example, information on a Web site may have been consolidated to only a few pages which caused foragers to

Table 100: Temporal Site-centric: Metric Statistics

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – SITE-PATCH						
RELDUR (in minutes)						
All	47	−2.10	3.07	−1.75	−11.98	4.45
Goal		0.27	1.58	0.23	−3.54	4.45
Non-goal		−4.46	2.27	−4.43	−11.98	0.18
RELPGS						
All	47	−0.15	1.25	0.00	−4.00	2.00
Goal		0.12	1.13	0.00	−3.00	2.00
Non-goal		−0.41	1.33	0.00	−4.00	2.00
STRICT INFORMATION SCENT						
RELUNQ						
All	47	−11.67%	15.72%	−8.33%	−50.00%	20.00%
Goal		−2.33%	11.03%	0.00%	−33.33%	20.00%
Non-goal		−21.02%	14.12%	−20.00%	−50.00%	8.93%
RELLNR						
All	47	−0.14	0.31	0.00	−1.00	0.46
Goal		0.00	0.09	0.00	−0.17	0.46
Non-goal		−0.28	0.38	0.00	−1.00	0.23

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

Relative clickstream linearity followed the same basic pattern as both the relative duration and number of pages viewed: values for goal sessions were positive while they were negative for non-goal sessions. On average, goal sessions had exactly the same value of clickstream linearity (0.00) as the previous goal sessions. Non-goal sessions had more than a quarter-of-a-point lower value (-0.28) for clickstream linearity than past goal sessions.

Table 101 lists the mean, standard deviation, median, minimum, and maximum values for the seven measures derived from patches and trails. The patches and trails were learned at the 0.05 significance level from prior goal and non-goal sessions. Each statistic is broken down for all, goal, and non-goal sessions. Each measure also lists the total number of Web sites that found patches or trails at any point during the processing procedure. The number of Web sites differ from the site-centric version (17 versus 14 patch Web sites and 15 versus 10 trail sites[11]) because of the multiple times patches and trails were mined at each Web site. For example, patches may have been found on a Web site when using 80% of goal sessions, but not when only 70% of goal sessions were used.

The first three patch measures (RELPTCMAX, RELPTCLAST, and RELPTCSUM) had average relative patch values of -0.17, -0.15, and -0.82 among all sessions, respectively. The relative patch values for RELPTCMAX and RELPTCLAST both had the same positive value (0.02). RELPTCMAX, however, was negative by almost a third of a point (-0.29). All three of the non-patch values shared negative values of -0.36, -0.31, and -1.36 for RELPTCMAX, RELPTCLAST, and RELPTCSUM, respectively.

Users spent, on average, 8.03 fewer seconds within patches relative to prior goal sessions. Current goal sessions spent 13.95 more seconds in patches relative to past goal sessions, whereas non-goal sessions spent 30.01 fewer seconds in patches.

Unlike the patch visitation measures, the trail following measures had negative values for all three groups of sessions. The average mean for RELTRLMAX, RELTRLLAST, and RELTRLSUM was -0.10, -0.09, and -0.22, respectively. All three measures for the goal sessions were also negative, but were close to having the same values as past goal sessions (-0.01 for RELTRLMAX and RELTRLLAST and -0.04 for RELTRLSUM). The non-goal sessions were much further away from zero than the goal sessions, with values ranging from -0.16 to -0.40.

switch back and forth between the pages.

[11] See table 58 in §7.2.1 for statistics on the site-centric version of the model.

Table 101: Temporal Site-centric: Metric Statistics (Significant – 0.05)

	N	Mean	St. Dev.	Median	Min	Max
INFORMATION PATCH – PAGE-PATCH						
RELPTCMAX						
All	17	−0.17	0.36	0.00	−1.30	0.30
Goal		0.02	0.08	0.00	−0.10	0.30
Non-goal		−0.36	0.43	−0.19	−1.30	0.00
RELPTCLAST						
All	17	−0.15	0.30	0.00	−1.02	0.30
Goal		0.02	0.07	0.00	−0.01	0.30
Non-goal		−0.31	0.35	−0.19	−1.02	0.00
RELPTCSUM						
All	17	−0.82	2.44	0.00	−11.44	1.62
Goal		−0.29	2.03	0.00	−7.99	1.62
Non-goal		−1.36	2.75	−0.43	−11.44	0.00
RELPTCDUR (in seconds)						
All	17	−8.03	39.61	−2.75	−118.00	102.75
Goal		13.95	33.62	5.88	−36.75	102.75
Non-goal		−30.01	32.85	−21.00	−118.00	4.63
RELAXED INFORMATION SCENT						
RELTRLMAX						
All	15	−0.10	0.27	0.00	−0.89	0.35
Goal		−0.01	0.17	0.00	−0.51	0.35
Non-goal		−0.18	0.33	0.00	−0.89	0.00
RELTRLLAST						
All	15	−0.09	0.24	0.00	−0.70	0.35
Goal		−0.01	0.17	0.00	−0.51	0.35
Non-goal		−0.16	0.28	0.00	−0.70	0.00
RELTRLSUM						
All	15	−0.22	0.92	0.00	−3.97	1.02
Goal		−0.04	0.80	0.00	−2.57	1.02
Non-goal		−0.40	1.03	0.00	−3.97	0.00

Note: all values are based on the median values from each Web site's goal and non-goal sessions.

8.2.2 Hypotheses Testing

Tables 102 and 103 present the results for the seven temporally-focused site-centric hypotheses. Table 102 provides results from the four hypotheses whose measure were not dependent on knowledge of mined patched and trails. Table 103 lists the results for the three hypotheses that relied on mined patches and trails.

The first two columns of each table list the hypothesis number and name of the metric being tested. The third and fourth columns list the total number of Web sites and the number of Web sites with a non-zero difference (i.e., $D_i \neq 0$), respectively. The total number of Web sites was used in the t-test, while only Web sites with non-zero differences were used for the Wilcoxon and sign tests. Columns five through seven list the t statistic, degrees of freedom (df), and p-value for the t-test. The eighth and ninth columns display the V statistic and p-value for the Wilcoxon test. The final two columns list the S statistic and p-value for the sign test[12].

[12] All three assumptions of the sign test were met. Therefore, the results from the sign test are focused on in the following paragraphs. Unlike the sign test, some assumptions of the Wilcoxon test (symmetry of D_is) and t-test (symmetry and normality of D_is) were not believed to have been met. Since the same data was used for both the temporal and non-temporal versions of the model, the same general unsymmetrical and non-normal distributions of D_is were expected. Thus, while results of the Wilcoxon test and t-test are provided in footnotes, the results of those tests should be interpreted with caution.

Table 102: Temporal Site-centric: Results

		N		T-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Ties	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – SITE-PATCH										
TSC1	RELDUR	47	47	13.87	46	$< 0.0001^{***}$	1,128	$< 0.0001^{***}$	47	$< 0.0001^{***}$
TSC2	RELPGS	47	30	2.22	46	0.0155^{**}	328	0.0243^{**}	20	0.0494^{**}
STRICT INFORMATION SCENT										
TSC7	RELUNQ	47	46	9.34	46	$< 0.0001^{***}$	1,049	$< 0.0001^{***}$	43	$< 0.0001^{***}$
TSC8	RELLNR	47	24	5.15	46	$< 0.0001^{***}$	295	$< 0.0001^{***}$	22	$< 0.0001^{***}$

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

Table 103: Temporal Site-centric: Results (Significant – 0.05)

		N		T-test			Wilcoxon		Sign Test	
Hyp.	Metric	Total	No Ties	t	df	p-Value	V	p-Value	S	p-Value
INFORMATION PATCH – PAGE-PATCH										
TSC5a	RELPTCMAX	17	10	3.74	16	0.0009^{***}	55	0.0010^{***}	10	0.0010^{***}
TSC5b	RELPTCLAST	17	10	3.95	16	0.0006^{***}	55	0.0010^{***}	10	0.0010^{***}
TSC5c	RELPTCSUM	17	10	3.32	16	0.0022^{***}	55	0.0010^{***}	10	0.0010^{***}
TSC6	RELPTCDUR	17	17	3.54	16	0.0014^{***}	142	0.0004^{***}	16	0.0001^{***}
RELAXED INFORMATION SCENT										
TSC9a	RELTRLMAX	15	6	2.13	14	0.0255^{*}	21	0.0156^{**}	6	0.0156^{**}
TSC9b	RELTRLLAST	15	5	2.21	14	0.0219^{*}	15	0.0313^{*}	5	0.0313^{*}
TSC9c	RELTRLSUM	15	6	2.37	14	0.0165^{**}	21	0.0156^{**}	6	0.0156^{**}

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01
Hypotheses SC5a-c and SC9a-c are each significant at $\frac{\alpha}{3}$ (e.g., $\frac{0.10}{3} = 0.0333$, $\frac{0.05}{3} = 0.0167$, and $\frac{0.01}{3} = 0.0033$).

TSC1 – RELDUR

The first hypothesis expected that goal achieving foragers would spend more time on a site, relative to prior goal sessions, than non-goal sessions would spend[13]. The results of the sign test supported hypothesis TSC1 at α = 0.01 (S = 47; p-value = < 0.0001)[14]. All 47 Web sites had a higher median relative duration amongst goal sessions than non-goal sessions. Relative to prior goal sessions, current goal sessions spent roughly 15 additional seconds on a site, while non-goal sessions spent almost five fewer minutes.

The results for this first hypothesis were identical between the two site-centric versions of the model (S = 47; p-value = < 0.0001 for both versions). For each of the versions, all of the tested Web sites had goal sessions with a higher median duration than non-goal sessions. Therefore, the use of duration, in either an absolute or relative manner, appears to be consistently useful in distinguishing between goal and non-goal sessions.

TSC2 – RELPGS

The second hypothesis also examined how foragers judged the value of a Web site, but did so by looking at the relative number of pages viewed. The hypothesis that goal sessions would have a higher relative median number of pages viewed than non-goal sessions was supported at α = 0.05 (S = 20; p-value = 0.0436)[15]. 20 out of the 30 non-tied Web sites had a higher median relative number of pages viewed for goal sessions versus non-goal sessions. On average, goal sessions viewed 0.12 more pages relative to past goal sessions, whereas non-goal sessions viewed 0.41 fewer pages.

The second hypothesis had practically identical results between the two site-centric versions of the model: static (S = 19; p-value = 0.0436) and temporal (S = 20; p-value = 0.0436). Roughly two-thirds of all the non-zero Web sites found a higher median number of pages viewed for goal sessions than non-goal sessions (67.86% of Web sites for static and 66.67% for temporal). This hypothesis also demonstrated that either an absolute or relative manner of determining number of pages viewed was useful in distinguishing between goal and non-goal sessions.

[13] A more in-depth discussion of each of the hypotheses may be found in §4.2.1 and §7.2.2.

[14] Hypothesis TSC1 was also significant at α = 0.01 for both the t-test (t = 13.87; df = 46; p-value = < 0.0001) and Wilcoxon test (V = 1,128; p-value = < 0.0001).

[15] Hypothesis TSC2 was also significant at α = 0.05 for both the t-test (t = 2.22; df = 46; p-value = 0.0155) and Wilcoxon test (V = 328; p-value = 0.0243).

TSC5 – RELPTCMAX, RELPTCLAST, and RELPTCSUM

The three sub-hypotheses of TSC5 (table 103) explored how the visitation of valuable patches could help explain goal achievement. All three sub-hypotheses were significant at α = 0.01 (S = 10; p-value = 0.0010 for all three measures)[16], supporting the hypothesized positive association of relative patch value and goal achievement. All 10 of the non-zero Web sites had goal sessions with higher relative patch visitation values than the non-goal sessions, for all three patch measures.

The results for hypothesis TSC5 were found to be significant at the same α level for both versions of the site-centric model: static (S = 9; p-value = 0.0020 for all three measures) and temporal (S = 10; p-value = 0.0010 for all three measures). Both versions of the model had all non-zero Web sites find a higher median patch value amongst goal rather than non-goal sessions. However, the temporal version had more Web sites find patches than the static version (17 versus 14 total Web sites). Thus, the temporal version was better able to utilize the available, but sparse, amount of data to learn patches. Furthermore, as the structure of a Web site may evolve, the temporal version's use of the most recent data would better reflect the changing nature of a site[17].

TSC6 – RELPTCDUR

Hypothesis TSC6 expected that mere visitation of valuable patches did not wholly indicate a forager obtained value from a patch. Thus, the hypothesis conjectured that relatively higher amounts of time within patches were associated with greater information gain and thus were more likely to achieve a goal. The results of the sign test supported the hypothesis at α = 0.01 (S = 16; p-value = 0.0001)[18], finding a higher relative duration within patches for goal sessions than non-goal sessions. 16 of the 17 non-zero Web sites with discovered patches had goal sessions that spent a higher relative duration of time within patches than non-goal sessions. On average, goal sessions spent almost 14 additional seconds within patches and non-goal sessions spent 30 seconds less.

Like many of the other hypotheses, the results between the two versions of the site-centric model for hypothesis TSC6 were also almost the same: static (S = 13; p-value = 0.0009) and temporal (S

[16] All three sub-hypotheses of hypothesis TSC5 were also significant at α = 0.01 for both the t-test (RELPTCMAX (t = 3.74; df = 16; p-value = 0.0009); RELPTCLAST (t = 3.95; df = 16; p-value = 0.0006); and RELPTCSUM (t = 3.32; df = 16; p-value = 0.0022)) and Wilcoxon test (V = 55; p-value = 0.0010 for all three measures).

[17] For a discussion of the limitations of the current incarnation of the temporal model refer to §9.1.

[18] Hypothesis TSC6 was also significant at α = 0.01 for both the t-test (t = 3.54; df = 16; p-value = 0.0014) and Wilcoxon test (V = 142; p-value = 0.0004).

= 16; p-value = 0.0001). Each version only had one Web site which found a higher median amount of time spent within patches for non-goal sessions than goal sessions (92.86% of positive Web sites for static and 94.12% for temporal). Therefore, the use of duration within patches appears useful in discriminating between groups of sessions, regardless of the measure being absolute or relative.

TSC7 – RELUNQ

The seventh hypothesis (TSC7) examined information scent in a strict manner where any inefficiency was viewed as poor indicators of scent. A positive association between the relative proportion of unique pages viewed and goal achievement was expected and supported at α = 0.01 (S = 43; p-value = < 0.0001)[19]. 43 of the 46 non-zero Web sites had goal sessions with a higher relative percentages of unique pages viewed than non-goal sessions. Both goal and non-goal sessions viewed a lower percentage of unique pages than past goal sessions (-2.33% versus -21.02%), but goal sessions still visited a greater proportion of unique pages than the non-goal sessions.

For this hypothesis, both the static and temporal versions of the site-centric model were supported at the same α level: static (S = 42; p-value = < 0.0001) and temporal (S = 43; p-value = < 0.0001). 95.45% and 93.48% of the non-zero static and temporal Web sites found goal sessions with a higher percentage of unique pages, respectively. Between the two versions, the unique percentage of pages viewed was equally successful in differentiating between the two groups of sessions.

TSC8 – RELLNR

The second hypothesis about strict information scent (TSC8) also examined information scent in a strict manner. However, overall scent was determined in a finer-grained manner by using the pages and the order in which those pages were visited. The belief was that less complex (i.e., more linear) clickstreams were indicative of higher levels of scent, and thus a greater likelihood of achieving a goal was expected and supported at α = 0.01 (S = 22; p-value = < 0.0001)[20]. 22 of the 24

[19] Hypothesis TSC7 was also significant at α = 0.01 for both the t-test (t = 9.34; df = 46; p-value = < 0.0001) and Wilcoxon test (V = 1,049; p-value = < 0.0001).

[20] Hypothesis TSC8 was also significant at α = 0.01 for both the t-test (t = 5.15; df = 46; p-value = < 0.0001) and Wilcoxon test (V = 295; p-value = < 0.0001).

non-zero Web sites had higher relative linear clickstream values for goal sessions compared to non-goal sessions, with goal sessions having, on average, the exact same clickstream complexity as prior goal sessions. Non-goal sessions were over a quarter of a point lower in clickstream complexity (-0.28) than past goal sessions.

The results of the static and temporal versions of the model were almost identical: static (S = 18; p-value = < 0.0001) and temporal (S = 22; p-value = < 0.0001). None of the non-zero static Web sites (0.00%) and only two of the temporal Web sites (8.33%) had any sites with non-goal sessions having a higher median clickstream complexity. Thus, like many of the other measures, both versions were equally capable of separating goal from non-goal sessions using the linearity of a user's session.

TSC9 – RELTRLMAX, RELTRLLAST, and RELTRLSUM

The final three sub-hypotheses of TSC9 (table 103) examined the efficacy that following valuable trails had in explaining goal achievement. Hypothesis TSC9a and TSC9c were both supported at α = 0.05 (S = 6; p-value = 0.0156 for both measures), while hypothesis TSC9b was only supported at α = 0.10 (S = 5; p-value = 0.0313)[21]. The difference in significance between the measures was due to sample size. Both RELTRLMAX and RELTRLSUM had six non-zero Web sites, while RELTRLLAST only had five (all of which supported the hypothesis in a positive direction). Thus, there were simply not enough Web sites for RELTRLLAST to reach significance at α = 0.05.

The results for hypothesis TSC9 were found to be significant at the same α level for all but one measure (RELTRLLAST) in the temporal version of the site-centric model: static (S = 6; p-value = 0.0156 for all three measures), and temporal (S = 6; p-value = 0.0156 for RELTRLMAX and RELTRLSUM and S = 5; p-value = 0.0313 for RELTRLLAST). Similar between the versions was all non-zero Web sites found higher median trail values within their goal sessions. However, just as

[21] Hypotheses TSC9a-c were significant at either α = 0.05 or 0.10, depending on the test. For the t-test, two of the three measures were significant at α = 0.10 (RELTRLMAX (t = 2.13; df = 14; p-value = 0.0255) and RELTRLLAST (t = 2.21; df = 14; p-value = 0.0219)), while the third was significant at α = 0.05 (RELTRLSUM (t = 2.37; df = 14; p-value = 0.0165)). For the Wilcoxon test, two of the measures were significant at α = 0.05 (V = 21; p-value = 0.0156 for RELTRLMAX and RELTRLSUM), while the third was only significant at α = 0.10 (V = 15; p-value = 0.0313). The less significant RELTRLMAX measure from the t-test may be due to the degree of difference between the goal and non-goal sessions. For example, both of the measures that were significant at 0.10 had less of an average difference between sessions (RELTRLMAX = -0.17; RELTRLLAST = -0.15) than the measure that was significant at 0.05 (RELTRLSUM = -0.35). The difference in significance between the measures of the Wilcoxon test was due to the same reason as found with the sign test: smaller sample size and thus less power to detect differences between the sessions.

with learning patches, the temporal version also had more Web sites find valuable trails than the static version (15 versus 10 total Web sites), highlighting the ability of the temporal version to use the extra available data to learn additional trails.

Summary of Results

Table 104 summarizes the results of the hypotheses testing. Of the 11 hypotheses and sub-hypotheses, seven were supported at α = 0.01, three at α = 0.05, and one at α = 0.10. The table also lists the results obtained from the static version of the site-centric CMIF.

Table 104: Temporal Site-centric: Hypotheses Results Summary

		Hypothesis Supported?	
Hyp.	Metric	Temporal	Static
INFORMATION PATCH – SITE-PATCH			
TSC1	RELDUR	Yes***	Yes***
TSC2	RELPGS	Yes**	Yes**
INFORMATION PATCH – PAGE-PATCH			
TSC5a	RELPTCMAX	Yes***	Yes***
TSC5b	RELPTCLAST	Yes***	Yes***
TSC5c	RELPTCSUM	Yes***	Yes***
TSC6	RELPTCDUR	Yes***	Yes***
RELAXED INFORMATION SCENT			
TSC7	RELUNQ	Yes***	Yes***
TSC8	RELLNR	Yes***	Yes***
RELAXED INFORMATION SCENT			
TSC9a	RELTRLMAX	Yes**	Yes**
TSC9b	RELTRLLAST	Yes*	Yes**
TSC9c	RELTRLSUM	Yes**	Yes**

*p $\leq$ 0.10; **p $\leq$ 0.05; ***p $\leq$ 0.01

8.3 Conclusion

Overall, the results between the two versions of the site-centric model did not differ in significant ways. Although the results were not significantly better, they were also not worse. Thus, the use of the temporal version provides additional evidence in the efficacy of the selected measures and in the ability of relative measures to distinguish between goal and non-goal sessions. In addition, the temporal version did see an increase in the number of Web sites which were able to learn patches and trails, although the significance of the results did not increase with the larger sample size.

At the surface, the lack of significantly better results than the static version would discourage the undertaking of the temporal model, especially given the computational cost and complexity associated with its methodology. However, the Web sites used to test the model may not have changed dramatically enough over the course of the data collection period to warrant the need for the temporal methodology. Warren et al. (1999) found within their limited examination of Web sites that "...the overall rate of change of a site increased with the size of the site" (pg. 182). Thus, the temporal version may be more appropriate for larger Web sites that are evolving at a faster rate than those seen within the site-centric dataset[22].

[22] On average, Web sites within the site-centric dataset were small with only 16.36 pages.

Chapter 9

Conclusion

This dissertation sought to explain goal achievement (i.e., choice behavior) at limited traffic long tail Web sites using Information Foraging Theory (IFT) (Pirolli, 2007; Pirolli and Card, 1999). The thesis of IFT was that individuals are driven by a metaphorical sense of smell that guides them through patches of information in their environment. Having a foundation in both psychology and ecology, IFT drew from both disciplines to explain the mechanisms and the resulting behavior of information foragers.

IFT used a production rule system from the psychological adaptive control of thought-rational (ACT-R) theory to describe the cognitive process of individuals foraging for information (Anderson et al., 2004). The rationalization of why a person would move from one area of their environment to another was explained according to the ecological patch model from optimal foraging theory (OFT) (Stephens and Krebs, 1986).

From ACT-R and OFT, the concepts of information scent and patches were defined for IFT. Information scent was the driving force behind why a person made a navigational selection amongst a group of competing options. As foragers were assumed to be rational, scent was a mechanism by which foragers could reduce their search costs by increasing their accuracy on which option lead to the information of value (Pirolli, 2007). An information patch was defined as an area of the search environment with similar information (e.g., single Web page, multiple Web pages, Web site) (Pirolli, 2007).

IFT was originally developed to be used in a “production rule” environment, where a user would perform an action when the conditions of a rule were met. However, the use of IFT in clickstream research required conceptualizing the ideas of IFT in a non-production rule environment. To meet such an end this dissertation asked three research questions regarding how to learn (1) information patches, (2) trails of scent, and finally (3) how to combine both concepts to create a Clickstream Model of Information Foraging (CMIF).

The first two research questions were similar in both concept and execution. In regards to patches, each user was free to define what a patch was as they saw fit. However, certain patterns of patches emerged on a Web site amongst those foragers with similar information goals. Likewise, scent trails were also defined by each user. When combined with other users, patterns from fragments of scent trails also emerged on a site between users with similar information goals. For the online firm, categorizing patches or trails as valuable to goal-achieving or non-goal-achieving foragers helped give an indication of the intent of users according to which patches or trails were visited or followed.

Research Question 1: *How can information patches be learned from a long tail Web site?*

Research Question 2: *How can information scent trails be learned from a long tail Web site?*

For research question 1 and 2, frequent itemsets and sequential patterns were learned on each Web site from goal and non-goal sessions to create contrast sets (Bay and Pazzani, 1999). Contrast sets which were able to significantly distinguish between the two groups of sessions at α = 0.05 were deemed valuable patches or trails. Once discovered, patches and trails were given a value according to how well the patch or trail distinguished between the goal and non-goal sessions.

In general, finding valuable patches and trails was successful on roughly a quarter of all tested Web sites (29.79% of sites for patches and 21.28% for trails). On those Web sites which did discover patches and trails, there were multiple instances of patches and trails being found (average of 11.93 patches and 4.70 trails per site).

The previous two research questions examined the concepts of information scent and patches individually. However, the real value of IFT was its ability to combine the search environment (i.e., patches) with the actions of a forager (i.e., scent). Thus the main focus of this dissertation and the final research question was on how these concepts could be combined using clickstream data to infer goal achievement.

Research Question 3: *How can information foraging theory and clickstream data be used to explain the achievement of a goal at a long tail Web site?*

Two versions of a clickstream model of information foraging were proposed which used clickstream metrics to represent the concepts of information scent and patches. In addition, the mod-

els also included measures which extended IFT. For example, hypotheses were introduced which tested the role of memory about a site and how patch value, specific to a group of foragers, could be used to predict goal achievement. The user-centric (UC) model exploited user-centric data (Padmanabhan et al., 2001) about a forager's entire browsing behavior to explain goal achievement at a long tail Web site. This model compared a forager's behavior across multiple Web sites. However, due to user-centric data being aggregated at the session level, the model lacked depth at individual Web sites.

In light of the rarity with which a user's entire clickstream over multiple sites is commonly available to an online firm, a site-centric (SC) version of the model employing site-centric data (Padmanabhan et al., 2001) was also developed. Having access to page-level data made the site-centric model capable of analyzing patches at all levels of analysis along with information scent at a Web site. However, since a forager's behavior across sites was unknown with site-centric data, the site-centric model compared a forager's behavior relative to an absolute value of zero[1].

The user-centric model proposed four hypotheses that examined the behavior of a forager within a site-patch (i.e., Web site). Three of the four hypotheses were supported at an α level of 0.01, while the fourth was not supported at any of the tested alpha levels. The site-centric model proposed the same four site-patch hypotheses as the user-centric model, plus the addition of two page-patch hypotheses, and three information scent hypotheses (nine hypotheses total). Five of the hypotheses were supported at an α level of 0.01, two at α = 0.05, and one at α = 0.10. The remaining hypothesis was found to be highly significant (α = 0.01) in the *opposite* direction of what was hypothesized.

Overall, both models were able to find measures which successfully distinguished between goal and non-goal sessions. Furthermore, the measures were grounded on a theoretical base that not only guided their selection (or creation), but also provided a reasoning for their existence that helped to explain why users behaved in the manners in which they did. In general, the two concepts of IFT were well supported using both versions of the clickstream model of information foraging.

The remainder of this chapter is organized as follows. First, the limitations of this research are discussed in §9.1. A discussion of the contributions of this dissertation are given in §9.2. Finally,

[1]Chapter 8 contains a temporal version of the site-centric model which compared each session relative to prior goal sessions.

§9.3 provides a brief overview of future research which expands upon this dissertation.

9.1 Limitations

As with any research, there were a number of limitations which should be recognized so that future research may improve upon this work. Listed below are nine limitations of this dissertation.

(1) Since IFT is a relatively new and not widely tested theory, basing this entire dissertation on its usage may be considered a limitation. However, even though the theory has not seen widespread usage like other theories commonly used in IS (e.g., Theory of Planned Behavior (Ajzen, 1991)), prior research has successfully used the theory. For example, elements of IFT have been used to inform the design of user-interfaces (Willett et al., 2007; Xie et al., 2006; Olston and Chi, 2003) and to help explain the browsing behavior of foragers (Lawrance et al., 2007; Galletta et al., 2006; Katz and Byrne, 2003). Furthermore, IFT is itself heavily based upon two theories that are well established within their respective disciplines: Optimal Foraging Theory (OFT) (Stephens and Krebs, 1986) and the Adaptive Control of Thought-Rational Theory (ACT-R) (Anderson et al., 2004). Therefore, while IFT is relatively new, its usefulness as a theory should not be discounted on that basis alone. Instead, this dissertation and other research like it are needed to determine, through evaluation, the worth of IFT.

(2) The prediction problem being examined between the user- and site-centric models were different. The site-centric model predicted if a goal would be achieved during the remainder of a session. To meet that task, only information that occurred *before* a form submission was used to calculate the measures and learn patches and trails. This forward-looking prediction was possible because the site-centric dataset contained page-level information, which allowed a session to be segmented such that only browsing behavior before the form submission was used. In contrast, the user-centric model predicted if a goal would have occurred given all information about a session (i.e., backward-looking prediction). The user-centric data was at the site-level and thus constrained the problem that could be analyzed. Since it was unknown where in the session a purchase took place, there was no reliable means with which to segment sessions.

    The use of all browsing behavior within the user-centric model introduced two limitations.

First, the measures reflected the browsing behavior of foragers before and after their purchase. While the data only allowed the first four measures to be tested, the change in information goal after the purchase may have introduced a greater amount of noise into some of the other measures (e.g., those dealing with page-patches and scent). The second limitation is that goal sessions by default would likely have higher number of pages viewed and session duration as a direct consequence of purchasing a product. For example, every goal session would have an increased number of pages viewed and session duration over non-goal sessions simply because they went through the checkout process. Thus, some of the differences seen between the measures of the first two hypotheses may be biased because all behavior from a session was used.

(3) Within the site-centric version of the clickstream model, the Web sites were assumed to remain relatively constant over the course of the data collection period. If the assumption of constant structure or content on a site was not met, then the browsing behavior of sessions may differ depending on when the sessions took place. For example, at one point in time goal sessions at a Web site may have visited 10 pages per session, on average. However, after reorganizing and streamlining the Web site, goal sessions then only viewed five pages on average. Comparing against an absolute value of zero would make distinguishing goal from non-goal sessions difficult because of the drastic change in browsing behavior.

    To combat this limitation a temporal version of the site-centric model was introduced in chapter 8. The temporal version compared all browsing behavior relative to all goal sessions which had taken place before the current session. Thus, the relative measures would be better able to reflect changes in the structure or content of a site. Comparing the results of the two versions of the site-centric model failed to find any large differences between the models, indicating the Web sites used in the site-centric dataset were mostly static. However, other datasets which contain Web sites which evolve at a much more rapid pace, may find better results using the temporal version of the model. Future research will more closely examine the affect time has on explaining goal achievement.

(4) The user-centric dataset contained Web sites of all popularity, but this dissertation was only interested in examining long tail Web sites. The limitation was a rigorous and quantifiable definition of what constituted a long tail Web site was not known. Thus, the 80/20 rule (Newman,

2005) was used to classify sites as either parts of the short head or long tail of a power law distribution. While the use of the 80/20 rule appears to be reasonable, future research should better explore how to define the long tail.

The user-centric dataset also restricted Web sites that were too far down the long tail. For example, sites with few achieved goals (< 50 purchases) were removed as they were suspected of being abandoned, too new, or representing failed business models. Due to their lack of traffic, these "very long tail" sites were considered too sparse to be usable for the intended analysis (e.g., mining may result in no or spurious patches and trails being found). While the selection of 50 goal sessions appears to be reasonable, the selection was specific to the user-centric dataset. Thus, future research may be better able to segment the long tail by defining generally-applicable rules.

(5) A common limitation faced when dealing with real-world datasets is the element of "noise" in the data. Although both datasets were preprocessed extensively, some elements of noise inevitably remained within the datasets. For example, within the site-centric dataset robots, spiders, and other automated programs may have been present in the data. To deal with these robots, the data provider had initially scrubbed the data for any self-identified robots. Then the outlier analysis was performed during preprocessing to remove any other out-of-place sessions.

The actual effect of such noise on the results of the model is unknown. However, it was believed the noise had a minimal impact because the results of both versions of the model generally came out as expected. Thus, the model demonstrates some level of robustness in the face of noisy data. Future work may be better able to quantify the impact noise has on the model by using more in-depth (e.g., categories of sites for the user-centric data) and focused data (e.g., browsing behavior from experimental participants).

(6) The determination of leaving and returning behavior within the same session differed between the two models, making comparisons of their results difficult. The user-centric model was stricter than the site-centric model in determining whether a session left and returned during a session. The user-centric model required a visitation of at least two pages at another e-commerce Web site, whereas the site-centric model counted visitations of any length at any

site. Thus, the site-centric model may have inadvertently introduced noise into the analysis by counting Web sites which were not related to the information goal of the forager at the site of interest[2]. Future research using more detailed user-centric datasets may be better able to determine if the distinction between the types of sites being left for makes a difference. For example, a site-centric dataset may be created from a detailed user-centric dataset[3] to determine if viewing more than one page at a similar type of Web site really matters when considering leaving and returning behavior.

(7) The site-centric version of the clickstream model used the first 70% of sessions to calculate patches and trails[4]. The rationale behind the usage of 70% was to have a large enough number of sessions to mine from in order to find valuable patches and trails, without introducing noise into the analysis by discovering spurious patches and trails (e.g., a patch that only one other session visited). Future research should perform a sensitivity analysis to determine if the results of the model change dramatically with different percentages.

(8) The temporal site-centric version of the model used the same set of sessions to perform two tasks. First, a set of goal and non-goal sessions were used to learn patches and trails. Second, the goal sessions from that same set of sessions were then used to calculate measures for patch visitation and trail following. The median value of those measures was then used to determine the relative value of patch visitation and trail following for the current session. The reason this approach was taken was due to the limited sample size at each Web site. Ideally, the mining of patches and trails should have used one group of sessions, while the calculation of measures should have used another. Future research that uses Web sites not so far down the long tail may be better capable of having independent groups of sessions accomplish each task.

(9) The path stratum measure was based on concepts from graph theory (McEneaney, 2001). When used to quantify the linearity of a user's clickstream two main limitations came to the surface. First, the path stratum measure would be much lower if the user started and ended

[2]Within the user-centric model, noise could have been further reduced by restricting e-commerce sites to those within the same product category as the target session. Unfortunately, the Web sites within the user-centric dataset were not categorized.

[3]See Padmanabhan et al. (2001) for an example of creating a site-centric dataset from a user-centric dataset.

[4]To be precise, the first 70% of goal sessions were used. In addition, all non-goal sessions which occurred before the last of the 70% of goal sessions were also used.

their session on the same page (e.g., the index page), as opposed to different pages (e.g., index and contact page). This is because the path the user took was a closed walk. Within the context of measuring scent, however; such a closed walk may not necessarily indicate such an extreme drop in scent. For example, a forager may return to the index page at the end of their session to make sure they investigated all links of interest.

The second limitation of the metric is that repeating sequential page views and multiple traversals of the same path are lost when transforming a clickstream to the converted distance matrix that is needed to calculate the measure. Since repeated behavior is lost, the overall scent of a user may be marked as high by the measure even in situations where multiple cycles occur within the clickstream. Given these limitations, future research may further explore if these situations unique to measuring information scent may be incorporated into the path stratum measure.

9.2 Contributions

In light of the limitations mentioned in the previous section, it is believed this dissertation still makes a number of worthwhile contributions. Listed below are the major contributions of this dissertation.

(1) First, this dissertation demonstrated how IFT could be used as a theoretical basis for clickstream research. Through the creation of two versions of a clickstream model of information foraging, the concepts of IFT were quantified outside of a production rule environment. In addition, the CMIF not only operationalized the core concepts of IFT, but also extended the theory by introducing memory, forager-independent valuation of patches and trails, along with refined definitions of scent (e.g., strict and relaxed scent). Once tested, many of the core aspects of IFT and the theoretical extensions introduced in this dissertation were supported. Thus, this dissertation not only demonstrated the ability of IFT to explain goal achievement, but it also introduced theoretical extensions which provided a more in-depth explanation of goal behavior.

(2) This dissertation also presented a methodology on how to learn patches and scent trails using not only significant, but also supported contrast sets. Measures were also created which quan-

tified a forager's visitation of patches and following of trails. The metrics measured the most valuable, last, and summation of all patches and trails that were visited or followed. For those Web sites within the CMIF that discovered patches and trails, the measures were capable of distinguishing goal from non-goal sessions according to a forager's visitation and following behavior.

(3) The third contribution was a methodology that detailed how to preprocess datasets with long tail Web sites. In particular, a separate user- and site-centric methodology was presented which highlighted the unique challenges associated with preprocessing each dataset. For example, a process was provided for the site-centric dataset about how to locate and select a single definable goal on Web sites which have more than one available goal.

(4) Finally, due to the presence of IFT guiding analysis, traditionally understudied long tail Web sites were able to be examined even in light of their sparse datasets.

9.3 Future Research

This dissertation was meant to provide a well-defined channel through which a stream of future research may flow. Thus, listed below are four future research projects that continue and extend upon the work in this dissertation.

(1) The first research project deals with attempting to answer the question "What is the long tail?". In this dissertation the long tail was defined as those Web sites which only accounted for 20% of achieved goals. A natural extension would be to more precisely define the separation between long tail and short-head Web sites. However, such a distinction may still be too simplistic in light of how much area the long tail portion of a curve may cover. Therefore, further segmentation within the long tail (e.g., the "very long tail") may also need to be defined.

In addition, there may be other means with which to define long tail Web sites, in general. For example, should sites be defined according to their total amount of traffic or by the number of goals achieved? If the goal being examined is purchases, can a site be a long tail Web site for one type of product, yet reside within the short-head for other product categories? If so, how do browsing patterns of foragers differ in regards to the long tailedness of the Web site's product categories?

The largest contributions of this research would be a clear definition of what "long tail" really means.

(2) The second research project is also a natural extension of this dissertation[5]: "How does the evolution of long tail Web sites affect browsing patterns?". The temporal version of the site-centric model provided an initial, yet somewhat simplistic, glimpse into a time-sensitive relative analysis. In essence, the temporal version used a window consisting of all previous sessions. However, including all previous sessions may be a liability on Web sites that commonly change, since "old" data would limit the ability of new patches and trails to be learned from the newly changed site. Thus this research project would examine how sliding windows may be defined to better meet the needs of long tail sites. For example, windows may be of a certain size (number or percentage), for a particular time period, of a size necessary to stabilize measures, or some combination of the three.

In addition, the burn-in period may also be defined such that measures are not calculated until patch and trail discovery has stabilized[6]. The use of stabilization may also have the added benefit of not "throwing" away extra banked sessions just because the bank had not met the prescribed number (or percentage) of sessions in it. Furthermore, the computationally expensive task of re-learning patches and trails may be restricted to only after those times of destabilization.

The largest contribution of this research would be a thorough analysis of how time impacts the analysis of foraging behavior on long tail Web sites. In addition, a methodology would be introduced that would make the most of the sparseness of data from long tail sites, while still allowing relative comparisons of foraging behavior.

(3) The third research project would provide a test of information foraging theory using a production rule system. In particular, IFT would be examined at long tail Web sites to determine how well the production rules, as specified by Pirolli (2007), are able to explain foraging behavior on long tail Web sites. In addition, production rules which take into account the theoretical

[5] A dataset which consists of Web sites that evolve at a more rapid pace than those seen in the site-centric dataset would be used.

[6] Measure calculation would also cease following Web site changes until patch and trail discovery had stabilized again.

extensions tested in this dissertation (e.g., memory, value of page-patches) would also be created and tested. The main contributions of this research would be two-fold. First, IFT would be tested in its original form on a sample of Web sites different from those sites used to create and test the theory. The second contribution would examine the ability and importance of the theoretical extensions outlined in this dissertation to explain goal achievement using IFT.

(4) The final research project would not be as direct of an extension of this dissertation as the other three projects, however; it would still employ IFT as a theoretical base to examine searching behavior. In particular, the purpose of this research piece would be to determine how search queries, used to arrive at a Web site, can predict the probability of a goal being achieved. The belief is that search queries are an observable manifestation of the information goal of a forager. Thus, information within a search query may provide clues into not only the goal of the forager, but also how well-defined the goal is. For example, assume one visitor submitted "flat-panel TV" for their search query, while another submitted "Sony Bravia 52". The first query appears to be more general in nature and thus may be more suited for browsing behavior that occurs during the information gathering stage. In contrast, the second query looks to be much more refined and pointing toward a specific product, which a forager may be interested in purchasing.

Semantic similarity, which is the likeness of concepts between two sets of words (Li et al., 2003), would be used to quantify the textual nature of search queries and then group similar search queries (and their resulting sessions) together. Clustering search queries, which are semantically similar to one another, may uncover groups of sessions which are more likely to achieve a goal during a session. The expected contributions of this research would be the introduction of semantic similarity to clickstream research and the creation of a methodology on how semantic similarity may be used to predict goal achievement.

References

Adamic, L.A. and Huberman, B.A. 2001. The Web's hidden order. *Communications of the ACM*. **44**(9) 55–60.

Ajzen, I. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*. **50**(2) 179–211.

Alonso, J.C., Alonso, J.A., Bautista, L.M., and Munoz-Pulido, R. 1995. Patch use in cranes: a field test of optimal foraging predictions. *Animal Behaviour*. **49**(5) 1367–1379.

Anderson, C. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, New York, NY.

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., and Qin, Y. 2004. An integrated theory of the mind. *Psychological Review*. **111**(4) 1036–1060.

Anderson, J.R., Budiu, R., and Reder, L.M. 2001. A theory of sentence memory as part of a general theory of memory. *Journal of Memory and Language*. **45** 337–367.

Anderson, J.R. and Milson, R. 1989. Human memory: An adaptive perspective. *Psychological Review*. **96**(4) 703–719.

Anderson, J.R. and Pirolli, P.L. 1984. Spread of activation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*. **10**(4) 791–799.

Awad, N.F., Jones, J.L., and Zhang, J. 2006. Does search matter? Using online clickstream data to examine the relationship between online search and purchase behavior. *Workshop on Information Systems and Economics*. Evanston, IL, 1–5.

Ayres, J., Flannick, J., Gehrke, J., and Yiu, T. 2002. Sequential pattern mining using a bitmap representation. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Edmonton, Alberta, Canada, 429–435.

Barnett, V. and Lewis, T. 1994. *Outliers in statistical data*. 3rd ed. John Wiley & Sons, Inc., West Sussex, England.

Bay, S.D. and Pazzani, M.J. 1999. Detecting change in categorical data: mining contrast sets. *The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, CA, 302–306.

Bhat, S., Bevans, M., and Sengupta, S. 2002. Measuring users' Web activity to evaluate and enhance advertising effectiveness. *Journal of Advertising*. **31**(3) 97–106.

Bloch, P.H., Sherrell, D.L., and Ridgway, N.M. 1986. Consumer search: An extended framework. *The Journal of Consumer Research*. **13**(1) 119–126.

Botafogo, R.A., Rivlin, E., and Shneiderman, B. 1992. Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*. **10**(2) 142–180.

Browne, G.J., Pitts, M.G., and Wetherbe, J.C. 2007. Cognitive stopping rules for terminating information search in online tasks. *MIS Quarterly*. **31**(1) 89–104.

Bucklin, R.E., Lattin, J.M., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J.D., Mela, C., Montgomery, A., and Steckel, J. 2002. Choice and the Internet: From clickstream to research stream. *Marketing Letters*. **13**(3) 245–258.

Bucklin, R.E. and Sismeiro, C. 2003. A model of Web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*. **40**(3) 249–267.

Burdick, D., Calimlim, M., and Gehrke, J. 2001. MAFIA: a maximal frequent itemset algorithm for transactional databases. *Proceedings of the 17th International Conference on Data Engineering*. Heidelberg, Germany, 443–452.

Byrne, M.D. 2001. ACT-R/PM and menu selection: applying a cognitive architecture to HCI. *International Journal of Human-Computer Studies*. **55**(1) 41–84.

Card, S.K., Pirolli, P., Wege, M.M.V.D., Morrison, J.B., Reeder, R.W., Schraedley, P.K., and Boshart, J. 2001. Information scent as a driver of Web behavior graphs: results of a protocol analysis method for Web usability. *Conference on Human Factors in Computing Systems*. Seattle, WA, 498–505.

Catledge, L.D. and Pitkow, J.E. 1995. Characterizing browsing strategies in the World-Wide Web. *Computer Network ISDN Systems*. **27**(6) 1065–1073.

Charnov, E.L. 1976. Optimal foraging, the marginal value theorem. *Theoretical Population Biology*. **9**(2) 129–136.

Charnov, E.L. and Orians, G.H. 1973. Optimal foraging: Some theoretical explorations.

Chatterjee, P., Hoffman, D.L., and Novak, T.P. 2003. Modeling the clickstream: Implications for Web-based advertising efforts. *Marketing Science*. **22**(4) 520–541.

Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., and Card, S.K. 1998. Visualizing the evolution of Web ecologies. *Proceedings of the SIGCHI conference on Human factors in computing systems*. Los Angeles, CA, 400–407.

Chi, E.H., Rosien, A., Supattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., and Cousins, S. 2003. The bloodhound project: Automating discovery of Web usability issues using the InfoScent Simulator. *Conference on Human Factors in Computing Systems*. Ft. Lauderdale, FL, 1–8.

Church, K.W. and Hanks, P. 1989. Word association norms, mutual information, and lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, B.C., 76–83.

Cochran, W.G. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics*. **10**(4) 417–451.

Collins, A.M. and Loftus, E.F. 1975. A spreading-activation theory of semantic processing. *Psychological Review*. **82**(6) 407–428.

comScore, Inc. 2005. comScore 2004 disaggregate dataset. URL `http://wrds.wharton.upenn.edu/ds/comscore/manuals/comScore_2004_Disaggregate_Dataset_WRDS.pdf`.

comScore, Inc. 2007a. 61 billion searches conducted worldwide in august. URL `http://www.comscore.com/press/release.asp?press=1802`.

comScore, Inc. 2007b. comScore 2006 Web behavior database. URL `http://wrds.wharton.upenn.edu/ds/comscore/manuals/comscore_wrds_2006.pdf`.

Conover, W. 1999. *Practical nonparametric statistics*. 3rd ed. John Wiley & Sons, Inc., New York, NY.

Dalgaard, P. 2008. *Introductory Statistics with R*. 2nd ed. Springer Science+Business Media, New York, NY.

Danaher, P.J., Mullarkey, G.W., and Essegaier, S. 2006. Factors affecting Web site visit duration : A cross-domain analysis. *Journal of Marketing Research*. **43**(2) 182–194.

Davison, B.D. 2000. Topical locality in the Web. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens, Greece, 272–279.

Engel, J., Blackwell, R., and Miniard, P. 1990. *Consumer Behavior*. 6th ed. The Dryden Press, Chicago, IL.

Ester, M., Kriegel, H., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. Portland, OR, 226–231.

Fader, P.S., Hardie, B.G.S., and Lee, K.L. 2005. RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*. **42**(4) 415–430.

Fielding, R., Gettys, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. 1999. Hypertext transfer protocol – HTTP/1.1. URL `http://www.ietf.org/rfc/rfc2616.txt`.

Fischer, J.E., Bachmann, L.M., and Jaeschke, R. 2003. A readers' guide to the interpretation of diagnostic test properties: Clinical example of sepsis. *Intensive Care Medicine*. **29**(7) 1043–1051.

Floyd, R.W. 1962. Algorithm 97: Shortest path. *Communications of the ACM*. **5**(6) 345.

Fu, W.T., Bothell, D., Douglass, S., Haimson, C., Sohn, M.H., and Anderson, J. 2006. Toward a real-time model-based training system. *Interacting with Computers*. **18**(6) 1215–1241.

Galletta, D., Henry, R., McCoy, S., and Polak, P. 2006. When the wait isn't so bad: The interacting effects of Website delay, familiarity, and breadth. *Information Systems Research*. **17**(1) 20–37.

Garshelis, D.L. 2007. *Brown Bear*. Oxford University Press, Oxford Reference Online. URL `http://www.oxfordreference.com/views/entry.html?subview=Main\&entry=t227.e19`.

Gourley, D. and Totty, B. 2002. *HTTP: The Definitive Guide*. O'Reilly.

Gray, W.D., Schoelles, M.J., and Sims, C.R. 2005. Adapting to the task environment: Explorations in expected value. *Cognitive Systems Research*. **6**(1) 27–40.

Hames, R.B. and Vickers, W.T. 1982. Optimal diet breadth theory as a model to explain variability in amazonian hunting. *American Ethnologist*. **9**(2) 358–378.

Harary, F. 1959. Status and contrastatus. *Sociometry*. **22**(1) 23–43.

Hawkes, K., Hill, K., and O'Connell, J.F. 1982. Why hunters gather: Optimal foraging and the Aché of eastern Paraguay. *American Ethnologist*. **9**(2) 379–398.

Helsel, D.R. and Hirsch, R.M. 1992. *Statistical Methods in Water Resources*. Elsevier Science Publishers B.V., Amsterdam, The Netherlands.

Hodge, V.J. and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review*. **22** 85–126.

Holling, C. 1959. Some characteristics of simple types of predation and parasitism. *The Canadian Entomologist*. **91**(7) 385–398.

Jaillet, H.F. 2003. Web metrics: Measuring patterns in online shopping. *Journal of Consumer Behaviour*. **2**(4) 369–381.

Janiszewski, C. 1998. The influence of display characteristics on visual exploratory search behavior. *The Journal of Consumer Research*. **25**(3) 290–301.

Johnson, E.J., Bellman, S., and Lohse, G.L. 2003. Cognitive lock-in and the power law of practice. *Journal of Marketing*. **67**(2) 62–75.

Johnson, E.J., Moe, W.W., Fader, P.S., Bellman, S., and Lohse, G.L. 2004. On the depth and dynamics of online search behavior. *Management Science*. **50**(3) 299–308.

Kalczynski, P.J., Senecal, S., and Nantel, J. 2006. Predicting on-line task completion with clickstream complexity measures: A graph-based approach. *International Journal of Electronic Commerce*. **10**(3) 121–141.

Katz, M.A. and Byrne, M.D. 2003. Effects of scent and breadth on use of site-specific search on e-commerce Web sites. *ACM Transactions on Computer-Human Interaction*. **10**(3) 198–220.

Kavassalis, P., Lelis, S., Rafea, M., and Haridi, S. 2004. What makes a Web site popular? *Communications of the ACM*. **47**(2) 50–55.

Kay, A. 2002. Applying optimal foraging theory to assess nutrient availability ratios for ants. *Ecology*. **83**(7) 1935–1944.

Lawrance, J., Bellamy, R., and Burnett, M. 2007. Scents in programs: Does information foraging theory apply to program maintenance? *IEEE Symposium on Visual Languages and Human-Centric Computing*. Coeur d' Alene, ID, 15–22.

Li, Y., Bandar, Z., and McLean, D. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*. **15**(4) 871–882.

MacArthur, R.H. and Pianka, E.R. 1966. On optimal use of a patchy environment. *The American Naturalist*. **100**(916) 603–609.

MacCracken, J.G. and Hansen, R.M. 1987. Coyote feeding strategies in southeastern idaho: Optimal foraging by an opportunistic predator? *The Journal of Wildlife Management*. **51**(2) 278–285.

McEneaney, J.E. 2001. Graphic and numerical methods to assess navigation in hypertext. *International Journal of Human-Computer Studies*. **55** 761–786.

Mendenhall, W. and Sincich, T. 2003. *A Second Course in Statistics: Regression Analysis*. 6th ed. Pearson Education, Inc., Upper Saddle River, NJ.

Miller, G.A. 1956. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*. **63** 81–97.

Moe, W.W. 2003. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*. **13**(1&2) 29–39.

Moe, W.W. and Fader, P.S. 2004a. Capturing evolving visit behavior in clickstream data. *Journal of Interactive Marketing*. **18**(1) 5–19.

Moe, W.W. and Fader, P.S. 2004b. Dynamic conversion behavior at e-commerce sites. *Management Science*. **50**(3) 326–335.

Montgomery, A.L., Li, S., Srinivasan, K., and Liechty, J.C. 2004. Modeling online browsing and path analysis using clickstream data. *Marketing Science*. **23**(4) 579–595.

Morrison, D.G. 1969. On the interpretation of discriminant analysis. *Journal of Marketing Research*. **6**(2) 156–163.

Mount, D.M. and Arya, S. 2006. ANN: a library for approximate nearest neighbor searching. URL `http://www.cs.umd.edu/~mount/ANN/`.

Nevitt, G.A. 2000. Olfactory foraging by antarctic procellariiform seabirds: Life at high reynolds numbers. *Biological Bulletin*. **198** 245–253.

Newman, M. 2005. Power laws, pareto distributions and zipf's law. *Contemporary Physics*. **46**(5) 323–351.

Nicholas, D., Huntington, P., Jamali, H.R., and Dobrowolski, T. 2007. Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the 'bouncer'. *Information Processing and Management*. **43**(4) 1085–1102.

Olston, C. and Chi, E.H. 2003. ScentTrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction*. **10**(3) 177–197.

Padmanabhan, B., Zheng, Z., and Kimbrough, S.O. 2001. Personalization from incomplete data: what you don't know can hurt. *The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, 154–163.

Park, Y.H. and Fader, P.S. 2004. Modeling browsing behavior at multiple Websites. *Marketing Science*. **23**(3) 280–303.

Pirolli, P.L. 2007. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, New York, NY.

Pirolli, P.L. and Card, S. 1999. Information foraging. *Psychological Review*. **106**(4) 643–675.

Pitkow, J. and Pirolli, P. 1997. Life, death, and lawfulness on the electronic frontier. *Proceedings of the SIGCHI conference on Human factors in computing systems*. Atlanta, GA, 383–390.

Pitts, M.G. and Browne, G.J. 2004. Stopping behavior of systems analysts during information requirements elicitation. *Journal of Management Information Systems*. **21**(1) 203–226.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Reader, W.R. and Payne, S.J. 2007. Allocating time across multiple texts: Sampling and satisficing. *Human-computer Interaction*. **22**(3) 263–298.

Ricca, F. and Tonella, P. 2000. Web site analysis: structure and evolution. *Proceedings of the 16th International Conference on Software Maintenance*. San Jose, CA, 76–86.

Ricca, F. and Tonella, P. 2001. Understanding and restructuring Web sites with ReWeb. *IEEE Multimedia*. **8**(2) 40–51.

Rode, C., Cosmides, L., Hell, W., and Tooby, J. 1999. When and why do people avoid unknown probabilities in decisions under uncertainty? testing some predictions from optimal foraging theory. *Cognition*. **72**(3) 269–304.

Rowley, J. 2000. Product search in e-shopping: a review and research propositions. *Journal of Consumer Marketing*. **17**(1) 20–35.

Sandstrom, P.E. 1994. An optimal foraging approach to information seeking and use. *Library Quarterly*. **64**(4) 414–449.

Satsangi, A. and Zaiane, O.R. 2007. Contrasting the contrast sets: An alternative approach. *Proceedings of the 11th International Database Engineering and Applications Symposium*. Banff, Alberta, Canada, 114–119.

Senecal, S., Kalczynski, P.J., and Nantel, J. 2005. Consumers' decision-making process and their online shopping behavior: a clickstream analysis. *Journal of Business Research*. **58**(11) 1599–1608.

Shaver, D. 1996. *The Next Step in Database Marketing*. John Wiley & Sons, Inc., New York, NY.

Simon, H.A. 1956. Rational choice and the structure of the environment. *Psychological Review*. **63**(2) 129–138.

Simon, H.A. 1974. How big is a chunk? *Science*. **183**(4124) 482–488.

Sismeiro, C. and Bucklin, R.E. 2004. Modeling purchase behavior at an e-commerce Web site: A task-completion approach. *Journal of Marketing Research*. **41**(3) 306–323.

Smith, E.A. 1983. Anthropological applications of optimal foraging theory: A critical review. *Current Anthropology*. **24**(5) 625–651.

Stephens, D.W. and Krebs, J.R. 1986. *Foraging Theory*. Princeton University Press.

Stewart, T.C. and West, R.L. 2007. Deconstructing and reconstructing act-r: Exploring the architectural space. *Cognitive Systems Research*. **8** 227–236.

Stone, B. and Jacobs, R. 2001. *Successful Direct Marketing Methods*. 7th ed. McGraw-Hill, New York, NY.

Streitberg, B. and Röhmel, J. 1986. Exact distributions for permutations and rank tests: An introduction to some recently published articles. *Statistical Software Newsletter*. **12**(1) 10–17.

Taatgen, N.A. and Anderson, J.R. 2002. Why do children learn to say "broke"? a model of learning the past tense without feedback. *Cognition*. **86**(2) 123–155.

Turney, P.D. 2001. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Proceedings of the Twelfth European Conference on Machine Learning*. Freiburg, Germany, 491–502.

Van den Poel, D. and Buckinx, W. 2005. Predicting online-purchasing behaviour. *European Journal of Operational Research*. **166**(2) 557–575.

Warren, P., Boldyreff, C., and Munro, M. 1999. The evolution of Websites. *Seventh International Workshop on Program Comprehension*. Pittsburgh, PA, 178–185.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*. **1**(6) 80–83.

Willett, W., Heer, J., and Agrawala, M. 2007. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*. **13**(6) 1129–1136.

Xie, X., Liu, H., Ma, W., and Zhang, H. 2006. Browsing large pictures under limited display sizes. *IEEE Transactions on Multimedia*. **8**(4) 707–715.

Yang, Q., Li, T., and Wang, K. 2004. Building association-rule based sequential classifiers for Web-document prediction. *Data Mining and Knowledge Discovery*. **8**(3) 253–273.

Yang, Y. and Padmanabhan, B. 2003. Segmenting customer transactions using a pattern-based clustering approach. *Proceedings of the Third IEEE International Conference on Data Mining*. Melbourne, FL, 411–418.

Zauberman, G. 2003. The intertemporal dynamics of consumer lock-in. *Journal of Consumer Research*. **30** 405–419.

Zhang, J., Fang, X., and Sheng, O.R.L. 2006. Online consumer search depth: Theories and new findings. *Journal of Management Information Systems*. **23**(3) 71–95.

About the Author

James A. McCart received a Bachelor of Science in Information Systems from Purdue University in 2002. In 2006, Mr. McCart earned a Master of Science in Management Information Systems from the University of South Florida.

While in the Ph.D. program at the University of South Florida, Mr. McCart was awarded a University Graduate Fellowship in 2005 and a College of Business Research Award in 2009. In addition, Mr. McCart was accepted in 2008 to Doctoral Consortiums for both the International Conference on Information Systems (ICIS) and America's Conference on Information Systems (AMCIS).