

# Golden Path Analyzer: Using Divide-and-Conquer to Cluster Web Clickstreams

Kamal Ali  
Vividence

1850 Gateway Drive, Suite 500  
San Mateo, CA 94404  
+1 650 645 5000

kamal3@yahoo.com

Steven P. Ketchpel  
Vividence

1850 Gateway Drive, Suite 500  
San Mateo, CA 94404  
+1 650 645 5177

stevek@vividence.com

## ABSTRACT

This paper describes a novel algorithm and deployed system Golden Path Analyzer (GPA) that analyzes clickstreams of people trying to complete the same task on a website. It finds the shortest, successful paths taken by users - 'golden paths' - and uses these as seeds for clickstream clusters. Other users are assigned to a cluster if their clickstream is a supersequence of the golden path. The advantages of this approach over prior work are that the resulting clusters are easily comprehended, they are few in number, correspond to semantically different strategies used by the users, and jointly partition all the clickstreams. GPA's key contribution over prior work in process funnels is that by not excluding users that make diversions from the golden path, GPA is able to assign more users to fewer clusters. Another key contribution is to use actual full clickstreams as cluster seeds to which supersequences of other users are added. Prior work on sequential variants of Agrawal *et al.*'s A Priori by contrast learned general but fragmentary clickstreams. GPA learns complete clickstreams that are based on actual user page transitions. GPA is particularly useful for site designers to improve processes such as shopping, returns and registration. Its analyses identify which web pages cause many users to deviate from a golden path, which links distract users and the percentage of users taking each golden path. GPA has demonstrated value on more than twenty client projects in diverse industries. It is implemented in Perl, Visual Basic and C#, runs in a Win2K/Intel environment and outputs a set of interlinked HTML pages.

## Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning

## General Terms

Algorithms.

## Keywords

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD 2003, Aug 2003, Washington DC.

Copyright 2000 ACM 1-58113-000-0/00/0000...\$5.00.

Web-mining, clustering, divide-and-conquer.

## 1. INTRODUCTION

Understanding how users are interacting with and navigating through web sites is a critical step in designing more effective, easy-to-use sites. A number of recent systems [2,4,5,10,11,12] attack this problem by clustering user clickstreams (e.g "searchers" versus "browsers"), characterizing which pages impeded or aided the tasks of the users, or determining which links distracted the users.

GPA is a clickstream clustering algorithm that assigns each user to one or more "golden paths," or instead, failure categories. A golden path is a sequence of pages, each reachable from the previous page, that leads from the initial page (such as the home page) to a page containing the information needed to solve the task, containing only pages relevant to the task. Each golden path corresponds to the actual path taken by at least one user and all the users are assigned the same, known task. A user's failure to achieve the task can usually be attributed to a small number of alternatives: "did not make a good faith effort", "failed to select a golden path", "fell off golden path G at page P", or "made it to final page but did not find where the answer was displayed". A fall-off is defined as failing to *immediately* transition to the next page in the golden path. Most successful users will have followed a golden path, but other succeeders may have merely guessed the correct answer or taken a rare path. Rare paths are not deemed to be golden, because they suggest uncommon knowledge on the part of the user, knowledge that should not be assumed of the typical user coming to the site. For example, asked to find the "least expensive digital scanner" on a site, someone who had recently purchased the item in question might be able to jump directly to the page, whereas the golden path would require using the "product search by price" page, a more involved process.

GPA was developed at Vividence, the leading provider of Customer Experience Management solutions for the web. It has been used in over twenty client engagements at Vividence. It has led to changes in navigational structure, link placement and messaging in those client sites.

The motivation to develop this algorithm stemmed from the fact that prior work on process funnels (a process like shopping has a funnel which tracks for each point in the process how many users followed along to that point) only learned clusters around sequences with overly strict inclusion criteria. If a user committed a diversion from the funnel, she was excluded from

the funnel. At the other extreme, algorithms like A Priori [1] learned very general but short and fragmentary subpaths.

Previous work on clustering clickstreams has taken the approach either of reducing this problem to the familiar one of clustering fixed-length vector data points (e.g. k-means clustering: [11]), finding longest common subsequences [2], finding commonly occurring (but potentially fragmentary) subsequences [1] or using mixtures of first order Markov models [5]. Comparison to GPA is covered fully in Section 6, 'Previous Work'some highlights are summarized here. One key differentiator for GPA (over [1,2,5,11]) is that each cluster has a seed, the golden path, which corresponds to a *full* clickstream of a user and represents one optimal way to complete the task. This golden path by itself usually provides a good explanation of the cluster (e.g. "searchers" versus "browsers") - explainability often being a weak point of clustering algorithms ([5,11]). The advantage of GPA over [1] for improving website processes is that each cluster description (the golden path) is a complete prescription for completing the task. Additionally, GPA has an advantage over prior process funnel work which excluded from the funnel any user who deviated from the funnel at all. GPA allows users to deviate (also called excursions or diversions) as long as they complete the rest of the path (potentially with other diversions). This means that a clickstream is assigned to the cluster of a golden path if it is a supersequence of the golden path.

The rest of the paper is organized as follows: Section 2 grounds the discussion by showing GPA running on a specific task, Section 3 rigorously specifies the algorithm, Section 4 describes the results it produces and Section 5 describes deployment and architecture. Section 6 compares it to other clickstream clustering algorithms and Section 7 offers suggestions for further research. Conclusions and contributions appear in Section 8.

## 2. AN EXAMPLE APPLICATION

Demonstrating GPA in the context of an actual site will make it easier to understand the inner workings. This section describes the findings of GPA for the Cars section of a major portal, dubbed FictitiousCarPortal.com. Visitors that come with the specific goal of finding a dealer quote on a specific model and option package have different ways of achieving their goal, along with many potential distractions and promising 'next steps' that lead them in the wrong direction. In order to discover the different paths visitors were taking, as well as where along those paths they were being distracted, Vividence sent a random sample of two hundred representative prospects to the home page, with the objective of finding a dealer quote. As part of standard Vividence methodology, measures were taken to ensure the set of people was representative of potential customers for that site. As they completed the task, their actions were tracked with Vividence Connector, a browser companion that records HTTP requests and other browser behavior. At the conclusion of the task, each user was asked several questions about the experience, including their ultimate satisfaction.

Two prominent links off the FictitiousCarPortal.com home page lead to different dealer quote processes. Figure 1 contains a simple schematic of the flow of these two different paths. These paths were automatically discovered by GPA. Although they seem trivial, of the 200 users, only 17 people followed the first

path exactly, step for step, and just 3 people the second exactly. In fact, the simplicity of these results is a strength of GPA since it is able to abstract away irrelevant diversions that individuals may have followed in their efforts to complete the task.

Further analysis identifies opportunities for site improvement. The GPA shows that many people "fall off" the golden path early in the process (at steps A1 and B1). The comments annotated to clickstreams from path 2 reveal people never receive a quote because there is no dealer in their area. Hence, coverage of the dealer network is a significant shortcoming of FictitiousCarPortal.com's site.

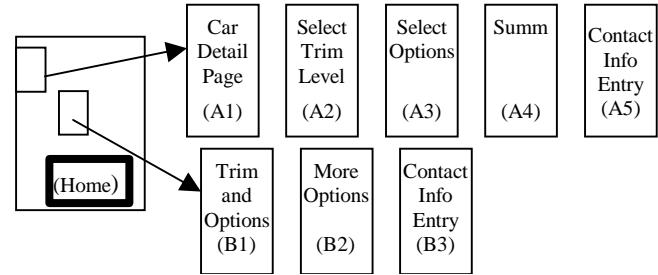


Figure 1: Golden paths discovered by GPA

A second question is whether it makes sense to have two different paths to achieve this objective. In this case, outputs from the GPA show that the two paths have comparable user satisfaction and success rates. Consequently, it seems that the two different flows are not detrimental to the user experience, and in general, having multiple entry points to key portions of the site will enable more people to find them. This GPA output suggests FictitiousCar-Portal.com might be better off having the two prominent links (with slightly different labels) share the same destination and process, removing the expense of maintaining the separate versions and the risk that they would report different prices.

With two distinct links from the home page into the quote process, it seems that FictitiousCarPortal.com is focused on ensuring its visitors find their way. However, the site includes three other links that distract many of the visitors at the home page. One in the left navigation bar, labeled "New Cars" along with a second link in the body of the page labeled "Research New Vehicles" both land on a page where the primary feature is a make/model selector that takes the visitor to a car detail page (A1 of Figure 1). The third main distracting link is one labeled "Buy a New Car," which leads to yet another make/model selector. This form submission takes people to the second golden path, where they re-join at the trim selection page (B1). In either case, the visitor has incurred an extraneous step that adds little to the customer experience. Perhaps FictitiousCarPortal.com prefers an additional page view and banner advertisement. In that case, they would be happy to know that none of the distractions resulting from these links was "fatal"; that is, people who followed them went on to successfully obtain their dealer quote, it just took an extra page view to do so.

Ensuring a clear flow between consecutive pages of a multi-phase process is one key to a good user experience with a high success rate. Another key part is making it clear when the

process is over, and what the final answer is. Here, FictitiousCarPortal.com fares well. In contrast, another major auto manufacturer was losing more than one-third of the people who wanted to submit a quote request, because midway through the process the visitor would receive a “Thank you for your interest in purchasing” message. Since they had entered their contact information in Step 1, many people mistakenly assumed that they were done with the process, not realizing that there were still three steps to go. In other engagements we have found the location of the desired information is easy to find, but the information itself is poorly communicated. People might follow the golden path to get to the target page, only to get the answer wrong when trying to interpret the text.

### 3. ALGORITHM

GPA analyzes the behavior of people who are trying to complete the same assigned task<sup>1</sup>, a set of well-defined instructions to attempt a common surfing task of commercial interest. The users are told to use the "ANSWER" button on the Vividence plug-in to their browser when they feel they have reached a page that has the answer to the objective question. An example objective question may be: “What is the price of the 2003 Toyota Corolla LE advertised at this site?”. Alternatively, they may use the "GIVE-UP" button and type in their reason for giving up. They may also at any point use a "COMMENT" button to type free text as to problems or frustrations they may be having. Augmenting clickstreams with such comments has proven to be immensely useful.

**Definitions:** In this paper, a *clickstream* is defined to be the sequence of pages actually traversed by a user. A *path* is an abstraction of a clickstream, so for example, the path  $\langle A, B, C \rangle$  may not have corresponded to any user's actual clickstream. *Sequence* is a generic term that may refer either to clickstreams or paths. A sequence of pages is represented as  $\langle a_1, \dots, a_n \rangle$ . Note that sequences are not sets so that  $\langle A, A, B \rangle$  is not the same as  $\langle A, B \rangle$  - the sequence  $\langle A, A, B \rangle$  may be generated by user refreshing page A and then moving to page B. GPA can and does learn golden paths with repeats of pages such as  $\langle A, B, C, B, D \rangle$  - the return to page B may be essential for some tasks in some web sites. Sequence  $s_1 = \langle a_1, \dots, a_n \rangle$  is said to be a *subsequence* of  $s_2 = \langle b_1, \dots, b_m \rangle$  if there is a mapping from indices in  $s_1$  to  $s_2$ ; that is if for  $1 \leq i \leq n$  there is an index  $i_j$  such that  $a_i = b_{i_j}$  and such that  $i_1 < i_2 < \dots < i_n$ . If  $s_1$  is a subsequence of  $s_2$  then by definition  $s_2$  is a *supersequence* of  $s_1$ . By this definition a sequence is its own supersequence. A sequence  $c$  is a *strict supersequence* of  $p$  if  $c$  is a supersequence of  $p$  and  $c$  is not equal to  $p$ . A sequence  $c$  is a *child* sequence of a *parent* sequence  $g$  if  $c$  is a strict supersequence of  $g$ . For example, the "child" sequence  $\langle X, A, D, B, E, C, F \rangle$  is a supersequence of the "parent" sequence  $\langle A, B, C \rangle$  because it contains all the elements of the parent sequence in order. The child sequence can have an arbitrary number of additional elements in between any adjacent elements in the parent sequence. In addition, it can have elements that precede the first element in the parent sequence and it may have elements that follow the last element of the parent sequence. A

<sup>1</sup> Non-compliant users that wander off the site are few and usually do not hinder discovery of golden paths.

user with clickstream  $c$  is said to have *followed a path*  $p$  if  $c$  is a supersequence of  $p$ . A path  $p$  is said to *cover* or *explain* a user  $u$  with path  $s$  if  $s$  is a supersequence of  $p$ .

**Inputs:** The key inputs to the GPA are the users' behavioral data and three parameters that control which user clickstreams are considered eligible candidates to be golden paths. Each user's behavioral data is a 3-tuple:  $\langle \text{clickstream}, \text{outcome}, \text{comments} \rangle$  where the clickstream is an ordered sequence of page visits (URLs) for a user, the outcome is one of  $\{\text{success}, \text{fail}, \text{giveup}\}$  and the optional comments are a concatenation of all the comments the user gave while doing this task. The outcome is typically decided by comparing the user's answer to a question (e.g. “How much does the Corolla cost?”) to the correct answer. Some GPA outputs depend upon a further input: each user's stated satisfaction with the process on a 7-point scale.

GPA requires the following three parameters as input:

**T:** at least  $T$  *succeeding* users must end at a page for that page to be considered a end point for golden paths. This threshold is used to exclude users who guess the right answer and hence are considered succeeders. If such a user ends on some arbitrary page it should not be considered a page that truly has the information necessary to answer the objective. Thus this parameter defines a set of pages a path must end on to be a candidate golden path.

**I:** to be a candidate golden path, the path must have been followed *exactly* by at least  $I$  succeeding users. This restriction prevents a single user who has “inside information” from skewing the results.

**C:** to be a candidate golden path, the path, or some supersequence of it, must have been taken by at least  $C$  succeeding users.

Taken together, the constraints imposed by parameters **T**, **I**, and **C** prevent guessers or users with inside information from being selected as golden paths. They also lead to learning of a few paths, each of which explains the clickstreams of many users. Rare paths are those that end at a legal end page and correspond to succeeding users but that fail to meet one or more of the **I**, **C** thresholds.

#### 3.1 Discovery algorithm

There are four stages in the core discovery algorithm:

1. Remove not-good-faith-effort (NGFE) users
2. Find legal end pages
3. Filter paths and sort paths
4. Divide and conquer

The system outputs the results as a set of interlinked HTML pages to allow browsing by the analyst in a variety of manners.

Stage 1 consists of removing users who have not given a good faith effort to complete the task. People who did not move beyond the start page are filtered out here. (This is actually an optional parameter that can be increased, requiring more pages.) In Stage 2, the set LEGAL-END-PAGES is defined as the set of pages that are terminated on by at least **T** succeeding users. Stage 3 forms the set of candidate golden paths by considering

only paths that start at the start page, end at LEGAL-END-PAGES, belong to a succeeding user, are followed by at least **I** succeeding users and each of which has at least **C** supersequences belonging to succeeding users. These candidate paths are then sorted: shortest to longest.

In the final *divide and conquer* [10] stage GPA is left with an ordered sequence of paths: GOLDEN-PATHS =  $\langle P_1..P_n \rangle$ . GPA then performs a linear scan of this sequence starting by considering  $P_1$ , which is a golden path. It cannot have any extraneous steps, because it is the shortest path that meets constraints established by parameters **T**, **I**, and **C**. At each step, GPA removes from GOLDEN-PATHS all supersequences of the path being considered. Finally it outputs the set GOLDEN-PATHS as an unordered set. This process normally yields one to four golden paths and explains about 60%-95% of the users. Note that it may be the case that the user is explained by more than one path. This can happen when the user completes a golden path but does not recognize the information on the final page, backtracks to the start and then completes a second golden path.

A crucial difference between GPA and previous work on clickstream funnels is that it allows users to make *diversions* or *excursions*. That is, for some contiguous pair of pages X, Y discovered as part of a golden path, a user does not need to *immediately* transition from X to Y; she can *eventually* transition from X to Y and still be counted as being explained by the golden path.

### 3.2 Additional Analyses

In addition to the value of identifying the set of golden paths, GPA further explains the user behavior through additional analyses utilizing the discovered golden paths:

1. **Falloff:** A user is said to fall off golden path G at some point P if instead of immediately going to the next element in the golden path, she goes somewhere else. GPA allows users to see which pages are responsible for the greatest number and percentage of falloffs. Falloffs are further split into three types: 1: a back move to a previous point (usually to the immediate predecessor) in the path, 2: go to another legitimate forward step corresponding to some other golden path (the *follow-set of page P*) 3: go to an arbitrary page. The follow-set of a page P is defined to be the union of all pages gleaned from all golden paths that can legitimately come after P. For example, in Figure 1 the follow-set of the HOME page is {A1, B1}. The idea is that from some pages, there is more than one legitimate next step. Therefore, falling off to another page of the follow-set is not a bad thing.
2. **Fatal falloff:** If the user fails to transition immediately to a page in the follow-set and if she never eventually returns to any of the pages in the follow set and she does not succeed at the task then that event is called a "fatal falloff". These are the worst errors, and should be remedied through changes in the site or page design.
3. **Distractor page:** When a user fails to choose a "next step" from the follow-set of the current page, GPA

allows one to easily see which pages people selected instead. This allows the web designer to see what is distracting the user from her main task. GPA lists the top distractor pages at each point in the process and tells how many users were distracted by each page.

4. **Non-recognizers:** Sometimes users will terminate on a target page (i.e.: one in the set LEGAL-END-PAGES) but still fail. Often the necessary information requires scrolling to find or else the wording is unclear.
5. **Overshooters:** These are non-succeeding users who completed one or more golden paths but who do not stop at the target page. This usually indicates that the target page does not make it clear that it is the final step in the process. Note that succeeders cannot qualify as overshooters because if they visited a target page, succeeded, but still went on, that may simply be because they wanted to do further explorations.

## 4. Architecture and Deployment

GPA currently exists in three implementations: a standalone Perl program which reads delimited flat-file input and outputs a set of interlinked HTML pages; a Visual Basic front-end and finally a C# version that reads from SQL Server databases with a Windows front end. At the end of a Vividence test, GPA is used by an analyst to recommend web site changes to the client.

Typically, for sample sizes of 400 to 800 users, GPA finishes in less than 10 seconds on a 2.4GHz Pentium 4 machine. The time complexity of finding non-good-faith-effort users and target pages are all linear in the number  $N$  of users. Sorting clickstreams by length is  $O(N \log N)$ . Finding the users covered by a golden path requires  $O(NL)$  steps where  $L$  is the clickstream length.

### 4.1 Challenge #1: Detecting Sessions

One of the significant challenges of interpreting clickstream data is ensuring that the data source is clean. Web logs, a common source used to infer the clickstream, have a number of disadvantages. First, an individual page may be composed of dozens of web hits, each of which generates an entry in the log. Second, not all of the user's navigations may be recorded by the web log. Especially in the case of a user re-visiting a page, the data may be served from the user's cache, omitting the entry in the web server log. Third, due to load-balancing, a given user's session may be fractured across several web servers. Fourth, mega-proxies such as AOL may serve the user from their own cache of the web page or use the same IP address for multiple users, making it hard to distinguish the requests of different people. Finally, detecting the end of a "session" is hard. Most work to date uses an arbitrary threshold of inactivity (15 to 30 minutes) as the end of the session.

An alternative approach to gathering the data, by collecting it at the browser, as each HTTP request is made, finesses all of these problems and provides a clean data stream. One remaining challenge is to filter out off-task activities (such as accessing a Hotmail account or checking an eBay bid). GPA uses data gathered by the Vividence Connector, which collects data at the client. Navigation events that are made in secondary windows

can be excluded from analysis if desired but in other cases may provide additional information on how the user is completing the task. Of course, the basic path discovery algorithm can be used with data from any source, however data collected from the browser directly (rather than a proxy or the server) is a better measure of the actual user experience.

## 4.2 Challenge #2: Page Aggregation

Given the dynamic nature of most sites on the web, the relationship between URL and the actual content is not perfectly reliable. It is common for a site to encode a session ID in the URL, meaning that each user has a unique URL, even though the content that they see is identical. In other cases, the content may differ, but not significantly enough to require distinguishing separate 'use cases'. For example, a dynamic page may list the car dealers within a user-specified zip code. For the purposes of determining a user's path, all of these pages should likely be considered equivalent, even though they have different URLs and content. This process of determining which pages should be treated as equivalent is called *page aggregation*.

The amount of aggregation that is desirable can be dependent on the particular analysis in question. In some cases, broad groups such as "stocks" or "news" can be manually identified and aggregated before clustering is performed. This is usually what has been done in previous work [4,5] that clusters at a coarse level. This is necessary because the raw pages, as represented by their URLs, are too variable from user to user - they may contain state about the user, or the time of day of the visit, or the specific stock being explored - details which may be immaterial to the analysis. Another source of variation is the definition of a "link". Two physically distinct links on a page may point to the same destination page: most work to date has not distinguished links at this level.

For GPA, a helper module does page aggregation as a preprocessing step. This page-aggregator operates on URLs such as `www.site.com/dir1/file.cgi?var1&var2=val2a&var3=val3a&...` and produces shortened URLs that then serve as page identifiers to the GPA algorithm. That is, GPA finds golden paths over some language of pages and the granularity of that language can be controlled by the analyst via the page-aggregator. The page-aggregator may decide to drop some variables such as the exact option package of a car or the zip code of the buyer. The current page-aggregator makes these decisions by applying heuristics to the distribution of values for each variable. If the variable has too many values then it is guessed to be a variable such as zip code, name or session ID and is dropped. Other simple rules cope with load balancing numbers in the URL, the optional "www" at the start of the URL, the optional presence of a substring such as `index.html`, `index.cgi` and so forth. The level of aggregation can also be overridden by an analyst who can make domain-dependent decisions on what variables to keep or drop. Note also that different analysts seeking different levels of detail may and do opt for different levels of aggregation.

## 5. RESULTS AND EVALUATION

The main outputs of GPA are HTML pages that:

1. Enumerate the set of discovered golden paths, annotated with numerous statistics on numbers of users

who followed the path to each point, fell off at that point and so forth. (See Figure 2 for a sample.)

2. Give an accounting of all the non-succeeding users in terms of whether they gave a good faith effort, whether they *selected*<sup>2</sup> a particular golden path and where they made their last (fatal) falloff from any golden path.
3. Give an accounting of all the succeeding users in terms of which golden path(s) they finished, and in some cases, of users that reached a legal end page without following one of the discovered golden paths. Such users are deemed to have followed a rare path that did not meet the parameter specifications  $\langle T, I, C \rangle$ .
4. Provide a very useful visualization (Figure 3) of the users paths sorted with respect to each golden path and secondarily with respect to the degree of completion within that path. Each golden path is assigned a color: say red for the first, blue for the second etc. The elements in that path are assigned increasing levels of saturation of the color so that the starting page is a very subtle red and the final page a heavy red. All users' paths are then sorted by length and each page that is a part of some golden path is colored with the appropriate hue and saturation so that it becomes easy to see what was their last point of contact with a given path or with any path. Pages that do not belong to any path are colored white. It also becomes easy to see when and where users switched from following one golden path to another or where they backed out of a golden path to restart from the starting page. Without this color scheme, it is difficult for humans to look at even a small set of clickstreams to comprehend what is happening.
5. Provide a table listing for each page in each path the set of top distractors for that page.

### 5.1 Visualizations

Figure 2 shows one of the main outputs of GPA: the two golden paths discovered (running down the page), together with many statistics pertaining to each step in the path. The figure shows idealized versions of the two golden paths that were actually discovered on a Vividence engagement for FictiousCarPortal.com and shown schematically in Figure 1. Additionally, GPA emphasizes (using bold font) statistics that are outside common ranges. So, for instance, in the "Immediate falloff" column (Column #9), the 60% number is emphasized indicating that that degree of "Immediate falloff" is higher than seen in prior engagements on most sites.

Each of the columns in Figure 2 provides a different piece of information about the golden path or its constituent pages. Column #1 is just a reference for which golden path the page level information pertains to. Column #2 is the identifier for the page (used also in the individual users' paths in Figure 3.) Column #3 counts the number of *followers*, those people that

---

<sup>2</sup> A user is said to have *selected* a golden path if she followed enough of the prefix of that path to disambiguate that path from any other golden path.

viewed each of the pages from the start page to the current page (though they may have viewed additional intervening pages). Column #4 counts the number of those followers that were successful, ultimately answering the objective correctly. The number of *visitors* (Column #5) is the number of users that viewed that page regardless of how they got there, so for each step in the golden path, the number of visitors is greater than or equal to the number of followers. Note that for the last step in golden path 1 (GP1), from Summary to Contact Info Form, the ratio of visitors (139) to followers (74) is much larger than it is for other steps in GP1. This is because the Contact Info Form step is also the last step in GP2. Therefore, the two golden paths share their starting and ending points but no other points in the middle.

Column #6 considers the number of users that left this golden path, but went directly to another golden path. Golden paths may share a common prefix, so falling off one may be a side-effect of choosing the other. In the first row of GP1, the entry in "#alt GP" indicates that 28 (out of the 198) visitors to "home" never eventually went to the second step "CarDetails" (GP1) but instead went to the next step in GP2.

Columns 7 through 9 distinguish three kinds of fall-off. The last, "Immediate falloff," is the easiest to understand. With respect to the transition from "Home" (HOME) to "CarDetails" (A1), a user is said to make an immediate falloff if s/he does not immediately go from HOME to A1. In this example, 60% of the visitors to HOME committed an "immediate falloff"; only 40% went directly to the next step in GP1. Note that some of those 60% may have eventually gone to A1, but they did not do so immediately. "Path falloff" means they never (i.e. not even "eventually") went to A1. They may have gone on to the other golden path but their interaction with this golden path is over - the figure shows that 39% of the users committed a "path falloff" at HOME and 21% more committed a "path falloff" at step A1. Once they have proceeded to step A2 ("selectTrim"), few of them commit a path falloff, indicating that they have become comfortable with the process from that point on. It may also indicate the absence of natural branching points in the web site at that part of the process. This is common for the start of a process. High falloff numbers in the middle indicate a serious problem. Note also that the "Path Falloff" associated with a step is merely the difference between the number of followers of the current and subsequent steps. Finally, "objective fatal falloff" means that the user never reached any of the pages in the follow-set of step HOME for any golden path and furthermore that they eventually failed or gave up on this objective<sup>3</sup>. The fact that they did not succeed also indicates that they also did not take a rare successful path. Their lack of success is typically ascribed to their "fatal falloff" at this point in the process: their failure probably points to problems on *this* page. The high number of immediate falloff (49%) at the last step "Contact Info Form" in GP1, would typically indicate that it was not clear to users that that was the end of the process. Here, however, many people

<sup>3</sup> The set of users committing immediate falloff is a superset of those committing a path falloff and that set in turn is a superset of those committing an objective falloff.

went further than was asked in the objective, actually submitting, rather than just filling out, an information request.

Column 10 indicates the eventual average path length for the set of followers at each step in the golden path. For example, for "CarDetails" (A1), the average path length is 9.9. That means that for the set of users that were following this path at least up to "CarDetails", that their average final clickstream length was 9.9. That is somewhat higher than the golden path length of 6, indicating the mean length of diversions. The figures in parentheses are the average path lengths for users who committed an immediate falloff at this point. This number is typically higher than it is for the people who didn't commit an immediate falloff. Note that in step "selectOptions" (A3), there is an unusually large difference between those making a falloff (14.8) versus those not (9.7) indicating that users that fell off at that point were likely to have a lengthy diversion, explained in this case by the opportunity to look at different option packages.

Finally, Column 11 indicates the average satisfaction for the set of users at each step, again contrasted with the satisfaction of people who fell off the path. For some web sites it is very interesting to note that the average satisfaction visitors completing different golden paths varies widely. Typically, falling off early in the process makes a larger difference in satisfaction than following off later in the path because the users that fell off at the end already have confidence that they are close to finishing.

<Fig 3: inserted approximately here>

A second visualization (Figure 3) drills down to the level of individual user clickstreams. It enables the analyst to see each step that a user took, and which of those steps overlapped with the discovered golden paths, and whether subsequent steps marked progress along a golden path or switching among golden paths, or many pages that were not a part of any golden path. For the FictitiousCarPortal.com site and vehicle pricing task, two paths were discovered. Although the GPA uses colors to distinguish among the golden paths, Figure 3 uses gray-scale shading to indicate depth of progress along a path, and the basic shape (square or oval) to indicate whether a page is a part of GP1 or GP2. For example, in Figure 3, user 1 follows GP1, and user 2 follows GP2. User 3 starts down GP2, but goes back to the home page before resuming his progress down the golden path. User 4 switches golden paths, taking the first step from the first path, but subsequently shifting over to GP2. User 5 is following along GP2, but visits a number of pages that are off-path while he does so. These pages are not necessarily "bad" or "disruptive", but they are not necessary to complete the task. Another benefit of this visualization is to join the behavioral data with text comments from the user giving insight into for example, *why* users were backtracking at some point. Without user comments, one could only infer or guess as to the reason. With the comment, though, it becomes clear that user 6 was trying to complete the objective, but was frustrated by a lack of available inventory.

## 5.2 Evaluation

To date, GPA has primarily been qualitatively evaluated by the quality of its results in terms of leading to client web-site modifications. It has led to navigational, structural and wording

changes on several client sites. For quantitative evaluation, we monitor the *coverage*, the fraction of the clickstreams that are explained by (covered by) at least one of the golden paths. If the learned paths do not cover a sufficient proportion of the testers, the analyst can relax the T,I,C thresholds or move to a coarser-grained page aggregation. The shortcoming of coverage is that if the model consisted of a separate golden path for each succeder, the explanatory power would be 100%, but the model would not generalize to other users' clickstreams. This extreme situation does not happen in practice for typical values of the  $\langle T,I,C \rangle$  parameters. In the future, for evaluation one could use cross-validation or a complexity penalty [e.g. 3] that favors models with fewer golden paths. In the current version, all of the small number (200 to 400) of clickstreams are used to build the model rather than setting some aside for validation. After the model was learned, human inspection was used to interpret the resulting golden paths and associated statistics. Future work could loop over different parameter combinations and page aggregations to obtain maximal coverage.

## 6. PREVIOUS WORK

Most previous work in the area of clickstream clustering addresses a slightly different problem than does GPA: explaining *all* of the visitors that came to a particular site, typically grouping them into different profiles based upon their objective or information need. GPA, in contrast, has a more focused data set and strives to determine which approaches the users took to answer the one assigned objective. GPA has an easier time of it because firstly it is dealing with client-side data (no sessionizing issues) and secondly the uncertainty about user objectives has been removed by our experimental methodology. Thus the system and analyst can focus on more semantic problems with the web site. In addition to the different goal, GPA uses a different clustering technique. Previous work has used K-means clustering over subsequences of paths [11], longest common subsequences [2] or mixtures of first-order Hidden Markov Models [5]. Not all of the related work described in this section uses clustering, some, such as that of Agrawal and Srikant [1] finds commonly occurring subsequences; other tools are exploratory data analysis assistants, such as WUM [14].

The k-means clustering algorithm ([8] has a good explanation and diagrams) requires a similarity measure for any two members of the universe. In [11], the similarity measure for two sequences  $s_1$  and  $s_2$  taken by users  $u_1$  and  $u_2$  relies upon the time spent on each page of the sequence. For  $sk$ , a subsequence of both  $s_1$  and  $s_2$ , let  $T_1(sk)$  be the time user  $u_1$  spent on  $sk$  and let  $T_2(sk)$  be the time user  $u_2$  spent on  $sk$ . The similarity of  $s_1$  to  $s_2$  is simply the sum of  $T_1(sk) * T_2(sk)$  over all subsequences shared in  $s_1$  and  $s_2$  that are less than a specified length,  $M$ . Having defined a measure of similarity, they then apply the K-means algorithm to cluster sequences. An attribute to the K-means algorithm here corresponds to a subsequence. The advantage of this approach is to recast the sequence clustering problem into the familiar realm of fixed dimensional clustering. However, one shortcoming of the approach is that very commonly taken subsequences are treated equally to subsequences that may only have been taken by two users. It also has a bias favoring long subsequences, since there are more subsequences of longer length and these longer sequences have more subsequences with

a greater total amount of time spent. GPA instead favors shorter sequences, which are deemed a more economical description of user behavior. A further disadvantage of the K-means technique is that it degrades as the number of dimensions increases and using a separate dimension for each subsequence leads to an explosion of dimensions.

Banerjee and Ghosh ([2]) also define the similarity between two sequences  $s_1$  and  $s_2$  using time spent on pages common to the sequence, though with a slightly different formulation. After placing pages into broad categories such as "sports" etc, they find the longest common subsequence, LCSS, between  $s_1$  and  $s_2$ . Let  $t_1(p)$  be the time spent by user 1 on some page  $p$  in LCSS and let  $t_2(p)$  be the time spent by user 2 on page  $p$ . They define the similarity between the sequences on page  $p$  to be the ratio  $\min(t_1(p), t_2(p)) / \max(t_1(p), t_2(p))$  which reaches a maximum value of 1 when  $t_1(p) = t_2(p)$ . They then define the similarity over the entire LCSS to be the page-wise average of these ratios over all pages  $p$  in the LCSS. Finally, they multiply the average by a scaling factor  $(t_1(LCSS)/T_1 * t_2(LCSS)/T_2)^{0.5}$ , where  $T_1$  is the time spent on all of sequence  $s_1$  and  $T_2$  is the time spent on sequence  $s_2$ . This facilitates comparison across LCSSs of different lengths giving higher weight to longer LCSSs. Having defined the similarity between any two sequences, they construct a fully-connected graph where each node is one of the sequences, and where the weight of the edge between two nodes is the similarity of the corresponding sequences. They then employ a minimal cutset algorithm to partition the graph into  $k$  subgraphs (the clusters) such that the sum of edge weights along cuts being made is minimized. The parameter  $k$  is specified by the user. Furthermore, if some sequence has no other sequence with at least some user-specified minimum similarity  $\theta$ , then the sequence is considered an outlier and is put into the "outliers" cluster. Their evaluation is done on 23,310 sessions (clickstreams) at [www.sulekha.com](http://www.sulekha.com), a community site. At  $\theta=.95$  (the value selected by the authors) about 17,000 (73%) of the 23,310 sessions were considered outliers. This work also differs from GPA in taking into account the amount of time spent on a page. GPA differs from the LCSS in that it permits diversions (the members of the subsequence do not need to be adjacent). GPA also prefers the shortest common subsequence that still explains the successful completion of the task. Finally, GPA designates one member of the cluster (the golden path) as being the best representative of the cluster in terms of explanatory power, whereas with Banerjee and Ghosh, all the cluster elements are equally important.

Large maximal sequences [1] are also an interesting comparison point to GPA. Agrawal and Srikant define a sequence  $S$  to be *large* with respect to some support level  $s \in [0,1]$  if the fraction of clickstreams in the observed data  $D$  that are supersequences of  $S$  is at least  $s$ . They define the sequence  $S$  to be *maximal* (for the given  $D$  and  $s$ ) if  $s$  is not a subsequence of any other large sequence. Informally, they seek sequences  $s$  that are as long (maximal) as possible whilst still being frequent (large) in the data  $D$ . This work is largely focused on efficiency -- finding such patterns in very large clickstream sets, using the *A priori* algorithm. The discovered patterns do not need to correspond to complete sequences taken by users, nor do they need to jointly

constitute a partitioning of the clickstream set. Most importantly, the found large maximal sequences do not even need to correspond to actual transitions made by the users. So, if the sequence  $\langle A, B \rangle$  is found it may simply mean that users that went to A eventually went to B but not that there need be any way to immediately get from A to B. This is touted as an advantage of the approach in being able to find remotely connected events (page visits). Thus although this work is able to find remote patterns which GPA may not find, the set of patterns found tends to be large and give fragmentary glimpses into the behavior of the users. It may be better suited for much larger datasets than those used for GPA and for free browse tasks where the users may have numerous objectives. However, the fact that the set of sequences found is large and fragmentary does not have the appeal of being able to fully partition the set of clickstreams into a small set of explanatory clusters and some left over rare paths as is done in GPA. In addition, it appears that following a small subsequence early in a process (e.g. detailed configuration of a car) is negatively correlated with following a small subsequence much later in the process (e.g. requesting help from car dealer). Both of these smaller subsequences would be found by *A priori* but the fact that they do not occur together can be seen in GPA by following the entire sequence of the user.

Clustering algorithms used have also included mixtures of first-order Markov models [5]. A first-order Markov model over some set of pages  $P$  is a  $|P| \times |P|$  matrix of transitions from any page in the set  $P$  to any other page. A mixture of such models allows for the fact that users' state may not be completely described by their presence on page  $P$  - rather, that each component (cluster) of the mixture models some different substate of those users that are at some page  $P$ . Cadez *et al.* [5] use the Expectation Maximization (EM) algorithm using out-of-sample log likelihood [8] to find the optimal number of clusters. EM differs from *k*-means in that each clickstream may belong with differing degrees of probability to each cluster. By contrast, *k*-means only allows for a clickstream to belong to a single cluster. Cadez *et al.*'s WebCANVAS system also provides for interactive exploration of the learned clusters, using colors to indicate page type (e.g. red for sports).

Work has also been done in defining a SQL-like language for exploring web sequences. WUM [14] is an interactive system that can display all sequences meeting criteria such as "sequences that visited a page A whose title is of the form "\*Corba\*", then eventually visited page B where  $\text{support}(B)/\text{support}(A) > 0.1$  (support here refers to the number of visitors to the page). WUM is mainly used by experienced analysts exploring the clickstream data. WUM does not itself find 'golden paths' for some success criterion but has more flexibility as a pure exploratory data analysis tool than does GPA. It could be used synergistically in a post-processing stage after GPA.

## 7. FUTURE WORK

Three areas of follow-on work would improve the robustness and value of GPA. First, a validation of the clustering would provide stronger theoretical underpinnings for a system that has proven useful in its practical application. The validation could take the form of running different clustering algorithms against the same

data set and comparing the results or comparing the results to expert-created "optimal" clusterings. Second, the GPA could be extended to a completely "hands off" system. Currently some human intervention is required for page aggregation (described in Section 4.2) and the selection of parameters  $\langle T, I, C \rangle$ :  $T$ : number of succeeding users needed to define a legal termination page,  $I$ : minimal number of identical succeeding users and  $C$ : the minimum number of children paths  $c$  that a path  $g$  must have in order to be a candidate golden path. Perhaps the addition of a complexity cost term [3] would enable the comparison of results produced under different settings of these input parameters, and ultimate selection of the best parameter settings and resultant model. However, the range of possible page aggregations makes this a daunting problem, especially since different people prefer different levels of aggregation depending on their analytic needs. Third, GPA could do more to highlight and explain anomalies in the data. The use of emphasis for "out of normal range" values as in Figure 2 is an example of the potential for this enhancement, but it would be better still to provide stronger explanations based upon the data and user comments for why these anomalies occur.

## 8. CONCLUSION

GPA has made the following contributions to the literature on clustering clickstreams: firstly, it has bridged the worlds of classification and web mining by applying the divide-and-conquer notion from work in decision trees [10] to aim for a complete picture of the set of clickstreams provided to it. Prior work on web mining consisted on the one hand of discovering full clickstreams (process funnels) that did not allow diversions and hence had little support. By searching in this limited model space, the possibility of finding longer paths with larger support is adversely impacted. On the other hand, algorithms like *A priori* find shorter subsequences with large support but do not provide linkage between subsequences followed early in the process with those found later in the process. Nor does *A priori* aim for a partitioning of users' clickstream behavior. GPA is also different because it is usually applied to a different problem: one of clustering clickstreams where there is a notion of success. Such situations are common in web site usage but little has been written on clustering for such objectives. GPA has also made practical contributions in improving navigational, structural and wording problems in commercial web sites.

## 9. ACKNOWLEDGMENTS

Thanks to colleagues at Vividence for valuable conversations in defining and applying the GPA. Special thanks to Mike Posner for building the C# implementation. Thanks to Andreas Weigend, Jeff Greenberg and Cliff Brunk for helpful feedback on this paper.

## 10. REFERENCES

- [1] Agrawal R. and Srikant R. Mining Sequential Patterns. Proc. IEEE ICDE, 3-14, March 1995
- [2] Banerjee, A. and Ghosh, J. Clickstream Clustering using Weighted Longest Common Subsequences, in Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining (Chicago IL, April 2001), 33-40.

- [3] Barron A., Rissanen J. and Yu B. The minimum description length principle in coding and modeling. *IEEE Trans. Information Theory*, vol 44 (1998), 2743-2760.
- [4] Berendt, B. (2002). Detail and context in Web usage mining: coarsening and visualizing sequences. In R. Kohavi, B.M. Masand, M. Spiliopoulou, & J. Srivastava (Eds.), *WEBKDD 2001 - Mining Web Log Data Across All Customer Touch Points* (pp. 1-24). Berlin etc.: Springer, LNAI 2356
- [5] Cadez I., Heckerman D., Meek C., Smyth P. and White S. Visualization of Navigation Patterns on a Web Site Using Model Based Clustering. In *Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, Massachusetts, August 2000. To appear. <http://citeseer.nj.nec.com/article/cadez00visualization.html>
- [6] Chen M.S, Park J. S. and Yu P. S. Efficient Data Mining for Path Traversal Patterns. In *Knowledge and Data Engineering*, 10(2): 209-221,1998
- [7] Cooley R., Tan P. N. and Srivastava J. Discovery of interesting usage patterns from web data, in: *Proc. of the Web Usage Analysis and User Profiling Workshop volume 1836 of Lecture Notes in Computer Science (2000) 163-182..*
- [8] Hand D., Mannila H and Smyth P. *Principles of Data Mining*. MIT Press, Cambridge MA, 2001.
- [9] Masand B. and Spiliopoulou M. (eds.). *Advances in Web Usage Mining and User Profiling: Proceedings of the WEBKDD' 99 Workshop*LNAI 1836. Springer Verlag, July 2000.
- [10] Quinlan J. R., *Induction of decision trees*, in *Machine Learning*, vol. 1, pp. 81--106, 1986
- [11] Shahabi C., Zarski A.M. and V. Shah J. Adibi. Knowledge discovery from users web-page navigation. In *Proc. 7th Intl Conf on Research Issues in Data Engg*, pages 20--29, 1997
- [12] Spiliopoulou M., Faulstich L. C. and Winkler K. A Data Miner analyzing the Navigational Behaviour of Web Users. In *Proc. of the Workshop on Machine Learning in User Modelling of the ACAI' 99*nt. Conf., Creta, Greece, July 1999
- [13] Spiliopoulou, M., Pohle, C. and Faulstich, L.C. Improving the effectiveness of a web site with web usage mining. In [Masand and Spiliopoulou, 2000], pages 139-159. 2000
- [14] Spiliopoulou M. and Faulstich L. C. WUM: A Web Utilization Miner. In *Workshop on the Web and Data Bases (WebDB98)* (1998), 109-115.

Figure 2: Golden Paths and Supporting Statistics

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Golden-path	Page	# followers	# succeeding followers	# visit	#alt GP	Obj. fatal falloff	Path falloff	Immediate falloff	Average Path length	Average Satisfaction
GP #1										
	home	198	136	198	<b>28</b>	9% (17)	<b>39% (77)</b>	<b>60% (119)</b>	8.2 (8.6)	4.6 (4.4)
(A1)	CarDetails	121	91	121	0	7% (8)	<b>21% (25)</b>	35% (42)	9.9 (11.1)	4.5 (4.1)
(A2)	selectTrim	96	76	103	0	(1)	5% (5)	8% (8)	9.8 (11.0)	4.7 (3.8)
(A3)	selectOptions	91	72	97	0	(4)	7% (6)	14% (13)	9.7 (14.8)	4.8 (3.5)
(A4)	Summary	85	70	93	0	(4)	13% (11)	21% (18)	9.3 (11.9)	4.9 (3.9)
(A5)	Contact Info Form	74	63	139	0	9% (7)	(0)	49% (36)	8.8 (9.8)	5.1 (4.7)
	Overshoots	9	0	-	-	-	-	-	9.1	5.2
	Did not recognize info	0	0	-	-	-	-	-	-	-
GP #2										
	Home	198	136	198	<b>51</b>	9% (17)	<b>59% (116)</b>	<b>60% (119)</b>	8.2 (8.6)	4.6 (4.4)

