INFOR, Vol. 50, No. 4, November 2012, pp. 186–194
ISSN 0315-5986 | EISSN 1916-0615

SOLO: A Linear Ordering Approach to Path Analysis of Web Site Traffic

Mark W. Lewis
School of Business, Missouri Western State University, 4525 Downs Drive, St. Joseph, Missouri, 64507, USA,
e-mail: mlewis14@missouriwestern.edu

Barbara Jo White
College of Business, Western Carolina University, 121 Forsyth, Cullowhee, North Carolina, 28723, USA,
e-mail: whiteb@email.wcu.edu

Abstract—Web usage mining analyzes web site traffic patterns in order to provide feedback on how they are being used. In this paper we present a new tool for web usage mining that is based on a linear ordering of the page transition matrix created from web server access logs. The ordering provides a measure allowing web pages to be categorized as origins, hubs or destinations according to their position in the ordering. It also provides a measure of the orderliness of web site traffic. This approach is applied to a university's web site traffic over time and results are discussed. Comparing web site traffic immediately after a major change to the web site design and then two years later, the traffic is more ordered. Results from the linear ordering approach are also compared to a Markov steady-state analysis of the page transition matrix. The mathematical formulation of the problem and the pseudocode used for its solution is presented and its application for bandwidth or size-limited devices is presented.

Keywords Web usage mining, linear ordering, network traffic analysis.

1. INTRODUCTION

Web usage mining, as described in Cooley (2003) and Spiliopoulou, et al. (2003) is studied in the context of various goals, such as improving individual user experience, balancing server loads and more accurately targeting markets. Specifically, web usage mining analyzes web site traffic patterns in order to provide feedback on how web sites are being used, what users are searching for and how the site might be improved.

Knowledge of how users are traversing a site allows web designers to adjust the web site to enhance the visitor experience by anticipating their actions (Montgomery, et al., 2004). For example, advertising effectiveness on the web is shown to be positively related to a user's "experience flow" as the user navigates a site (Huizingh and Hoekstra, 2003). They show that related advertisements along a path are more effective than a single one. In the complex interconnectedness of the web and of websites, providing common paths through the maze, based on a starting position and leading to popular end points, has great value.

In this paper, we propose a method, Sequencing in Optimal Linear Order (SOLO), of generating these paths using linear ordering (LO). A linear ordering is distinguished from a hierarchical ordering created by human editors or from indices of terms created by web crawlers in that a linear ordering creates a sequentially ordered list of web pages best fitting the traffic observed. To our knowledge, this technique has never been applied to web traffic. In this paper, we apply LO to a university web site using its server logs from two spring semesters, one immediately after a major site redesign and the second analysis done two years later.

When users visit a web site, information on the pages they visit is stored in the server's access logs. These logs can be analyzed to determine simple metrics such as most visited pages, most time spent on a page and busiest times of the day. They can also be used for more sophisticated path analysis using a transition matrix containing the number of traversals made from one web page to another by users, as identified by the IP addresses of their computers. The linear ordering of this page transition matrix generates a best-fit order in which pages are visited by all recorded users, from entry into the site to navigation around the site and then exit from it.

A linear ordering allows web pages to be categorized as origins, hubs, and destinations, corresponding to beginning, middle and end points respectively. Thus LO creates a new method for, and interpretation of, the rank ordering of web pages, where rank is the position in the linear order. Aggregating LO

Received April 2012; Revised August 2013; Accepted August 2013

solutions over time allows web site designers to analyze traffic flow and help direct it by adding useful links. Web pages that are consistently ranked in the same categories allow highly summarized site maps to be created for use on devices with small display real estate or limited bandwidth, for example.

The linear ordering metric (LOM), derived from the objective function value of SOLO's optimally ordered transition matrix, allows us to introduce a measure of the randomness of traffic flow. A high LOM indicates that most traffic is starting from and heading towards the same web pages and visiting the same intermediary pages along the way.

In this paper, we first describe the SOLO linear ordering approach in Section 2 and then provide a mathematical problem formulation and solution in Section 3 that includes an illustrative example and implementation code. Section 4 presents the results from an analysis of web server logs over a two year period.

2. DESCRIPTION OF THE SOLO LINEAR ORDERING APPROACH

Our approach to the ordering and categorization of web pages, SOLO, differs from commonly-used methods of page ranking and web site summary statistics in several ways. The original PageRank (Brin and Page, 1998) system developed by the founders of the Google search engine, used the number of hyperlinks that refer to a page (i.e. number of citations) to help determine that page's rank amongst all pages registered as containing the text of a query. For example, suppose that a user requests information from a search engine by using word *W* and suppose that page *P* contains word *W*, then *P*'s rank within the list of all pages containing *W* is based on the number of other pages that have links to *P*. Thus, the links citing another page act as endorsements in helping to determine which pages are deemed most relevant and therefore placed at the top of the list to be viewed by the user.

In SOLO, however, it is the interaction between pages, in other words the *transitions*, or the actual movement from one page to another, that determines the ranking of a page. The page transitions document active interest in the requested page and are expected to provide more accurate rankings than counting hyperlink referrals that no one may actually follow.

Page transitions are similar to click-throughs, a term generally associated with the number of user clicks generated by an advertisement. Click-through rates (number of ad clicks divided by number of exposures) are used by internet marketing firms to gauge advertising effectiveness. Besides click-through data, other various commercial products help analyze web site traffic by providing simple summary statistics of page view volume, ratios of page views to number of visitors, average time on site, bounce rate (percentage of people leaving after viewing a single page), or various tree structures for visually describing traffic between pages (Berkhin et al., 2001).

In contrast to these summary statistics, SOLO provides a linear ordering which allows for visualization of the general traffic flow. Figure 1 illustrates this concept using net traffic flow between a linear ordering of selected web pages.

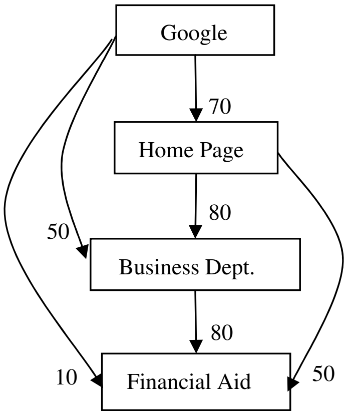

Figure 1. Illustrative Linear Ordering Moving from Google towards Financial Aid

The optimal linear order produced by SOLO, however, is not a strict hierarchical precedence, requiring one page be visited before another. Rather, it is an arrangement of pages in an order providing the greatest agreement with the traversals documented in the page transition matrix. Campos, et al. (2001) and Reinelt (1985) discuss how the LO problem can also be viewed as finding the acyclic tournament with maximal arc sum weight.

The optimal linear order produced by SOLO also differs from a site map (see Figure 2), which shows the hierarchical structure of pages in directories in the web site and is provided to help search engines and web crawlers index and report the site's pages correctly. Site maps also serve as navigational aids to users.

In this paper, using SOLO, we aggregate all web page requests below the first level in a hierarchical structure. Thus, transitions between pages located in different directories (branches) were analyzed in this research, but not transitions between the pages

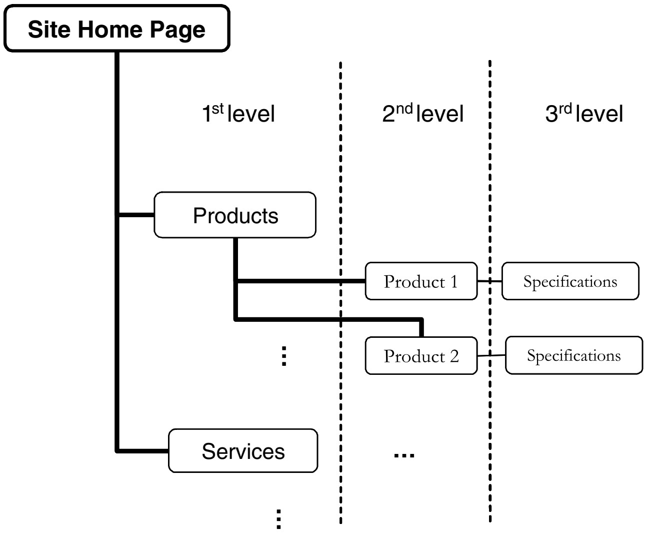

Figure 2. Example Hierarchical Site Map with Products and Services

ISSN 0315-5986 | EISSN 1916-0615
DOI 10.3138/infor.50.4.186

stored at lower levels of a common first level directory. Further, SOLO utilizes server-side, anonymous, non-intrusive server logs to collect and analyze web user traffic, as opposed to tracking via cookies, spyware or login information. From these server logs, large amounts of page request data are cleansed, user paths determined, and a page transition matrix created. This matrix is used to create the corresponding LO problem.

The linear ordering of a page transition matrix maximizes the sum of the elements in the upper triangular portion of the matrix by reordering the rows and columns while preserving the page transition data. LOM is the ratio of the sum of the upper elements to the sum of all elements in the page transition matrix. The optimal linear ordering is a "best fit" to the traffic data aggregated in the transition matrix. Although solving LO problems to optimality is NP-hard, according to Chanas and Kobylanski (1996), very good solutions can be quickly found using an efficient scatter search metaheuristic available in Laguna and Martí (2003). A perfect ordering that exactly fits the page navigations of all visitors would be rare, unless the pages analyzed were extremely structured, such as a sequence of pages with forms to fill out when making an e-payment for a product.

Generally, the LOM differs depending on the goals of the web users. For example, web users can be modeled as random or intentional (sometimes referred to as the hike or hunt metaphor). Intentional users are searching for specific information while random users are simply browsing with no specific direction, similar to a random walk. The LOM of web site traffic with intentional users should be higher (indicating more order) than web site traffic with random users. This research does not presuppose either of these two models. The web page traversals of all users of the university web site were aggregated into one page transition matrix for each day, with no regard for user classification. Since a university web site is typically a diverse collection of web pages, the page transition matrices in this research are not expected to have a high degree of order (high LOM). However, certain pages are expected to be consistently ranked as either origins, hubs or destinations.

The web page structure of the internet is described as being composed of *hubs* and *authorities* in Ding, et al. (2004). *Authorities* are the informational pages that a user is searching for and *hubs* are pages containing lists of links to authority pages. In this paper, we expand on this concept and include the category of *origins* as pages from which traffic originates, usually by a browser bookmark or a link from an external source such as a search engine or advertisement. Origins appear towards the beginning of a linear ordering. We use the word *destination* instead of authority to describe pages that users are moving towards, that is, they are found towards the end of a linear ordering. Pages occurring in the middle of an LO are categorized as hubs.

Zheng, et al. (2003) noted that slight changes in the techniques employed to aggregate user data can create large changes in the conclusions drawn about them. They used client-side monitoring of web activity to obtain data about a user's web activity. Client-side monitoring requires software to monitor and report user web activity, similar to spyware or the use of tracking cookies by some advertisers for tracking customer browsing across different web domains. In this paper, we use the data already being collected in the server logs, at the server-side, and *do not* attempt to aggregate data by user type, for example as inferred from the IP address of each user. However, analyzing differences in linear orderings between categories of users would be a topic for future research; for example, based on their linear orderings, are different groups of users categorized as intentional or random browsers?

In our research data, the daily server log files would have from 50,000 to over 500,000 page requests for over 4,000 unique URLs (Uniform Resource Locator). An example of the type of data collected by a server is shown in Figure 3.

This data can be sorted by user and time, so that the number of requests from one page to another can be tallied, creating a transition matrix for the LO problem and which can also be used for Markov analysis. No personally identifiable information referring to the actual user of the computer is collected. IP addresses associated with web crawlers were removed from the data analyzed.

3. SOLO PROBLEM FORMULATION AND SOLUTION METHOD

Given an $n \times n$ matrix of weights $C = \{c_{ij}\}$, we wish to find the permutation of rows and columns of C such that the sum of the weights of the upper triangular matrix are maximized. Letting $x_{ij} = 1$ if item i precedes item j in the ordering, the standard binary integer programming model is:

$$\text{Max} \sum_{i<j}^{n} c_{ij}x_{ij} + \sum_{j<i}^{n} c_{ij}(1 - x_{ji}) \qquad (1)$$

```
#Software:  Microsoft
#Version:   1
#Date:      3/1/2006
Date        time       c-ip              cs(Referer)
11/1/2005   6:00:00    12.123.123.123    http://www.missouriwestern.edu/Business/index.html
11/1/2005   6:00:02    12.123.123.123    http://www.missouriwestern.edu/Business/faculty.html
11/1/2005   6:01:12    12.123.123.123    http://www.missouriwestern.edu/Admissions/index.asp
```

Figure 3. Example Server Log Data Used to Create Paths and Transition Matrix

$$\text{s.t} \quad x_{ij} + x_{jk} - x_{ik} \leq 1 \qquad \forall\ (i,j,k):\ i < j < k \tag{2}$$

$$x_{ij} + x_{jk} - x_{ik} \geq 0 \qquad \forall\ (i,j,k):\ i < j < k \tag{3}$$

$$x_{ij} \in \{0,1\} \qquad \forall\ (i,j):\ i < j \tag{4}$$

Alternatively, given an $n \times n$ matrix $C = \{c_{ij}\}$, where c_{ij} denotes the total number of users on page i that request page j, the LO problem seeks to maximize the sum of the requests in the upper triangular portion of the matrix by manipulating the rows and columns of C while preserving the page request data c_{ij}. To accomplish this manipulation of rows and columns, define a permutation of rows i of C as the array $p(i) = [i_1, i_2, \ldots, i_n]$ with the initial ordering $p = [1, 2, 3, \ldots, n]$, i.e. $p(1) = 1, p(2) = 2$, $p(n) = n$. The linear ordering problem reorders the elements of p to maximize

$$\text{Upper}_C(p) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{p(i)p(j)}$$

where $c_{p(i)p(j)}$ is the number of requests from page *p(i)* to *p(j)*. Swapping elements in the array p generates a new ordering and a new $\text{Upper}_C(p)$ value. We define the Linear Ordering Metric (LOM) as the upper triangular sum of C divided by the sum of all requests in C. LOM is an easy metric to understand and it allows for comparisons of orderliness.

$$\text{LOM}(p) = \text{Upper}_C(p) / \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}$$

Because LO is NP-Hard, large problems are not amenable to exact solutions, therefore we used the Greedy Randomized Adaptive Search Procedure (GRASP) C code developed and made publicly available by Laguna and Martí (2003) to quickly find an excellent ordering (although not proven optimal). GRASP works very well with problems such as LO which involve permutations of matrices. In Lewis, et al. (2009), the modeling and solution of LO was compared between GRASP, Cplex (using the LO integer programming model) and a multi-start tabu search with path relinking approach, with GRASP providing high quality solutions very quickly. For example, transition matrices of size 200×200 were solved by GRASP in 0.7 seconds and for the average 100 sites ordered per day in this paper, linear orderings were created in 0.1 seconds.

3.1 Illustrative example

A small example illustrating sequencing in optimal linear ordering is now provided. From the server logs of a university, information related to traversals involving four high level pages has been extracted to form the transition matrix shown in Figure 4a. The matrix indicates that there were 120 requests from web pages associated with the Business School to pages associated with Financial Aid, and 40 transitions from Financial Aid to Business.

		Business	Financial Aid	Google	Home Page		
		1	2	3	4		
Business	1	0	120	10	30	330	← Upper Triangular Sum
Financial Aid	2	40	0	20	60	390	← Lower Triangular Sum
Google	3	80	50	0	90		
Home Page	4	110	90	20	0	46%	LOM

(a)

		3	2	1	4		
Google	3	0	50	80	90	350	← Upper Triangular Sum
Financial Aid	2	20	0	40	60	370	← Lower Triangular Sum
Business	1	10	120	0	30		
Home Page	4	20	90	110	0	49%	LOM

(b)

		3	4	1	2		
Google	3	0	90	80	50	540	← Upper Triangular Sum
Home Page	4	20	0	110	90	180	← Lower Triangular Sum
Business	1	10	30	0	120		
Financial Aid	2	20	60	40	0	75%	LOM

(c)

		3	4	2	1		
Google	3	0	90	50	80	460	← Upper Triangular Sum
Home Page	4	20	0	90	110	260	← Lower Triangular Sum
Financial Aid	2	20	60	0	40		
Business	1	10	30	120	0	64%	LOM

(d)

Figure 4. Swapping Matrix Elements in Small SOLO Linear Ordering Problem. (a) Site order $p(i) = [1,2,3,4]$. (b) Swap Google and Business, site order $p(i) = [3,2,1,4]$. (c) Swap Financial Aid and Home Page, optimal site order $p(i) = [3,4,1,2]$. (d) Swap Financial Aid and Business site order $p(i) = [3,4,21]$

With the ordering $p(i) = [1,2,3,4]$ as indicated in Figure 4a, Business is first in the ranking, second is Financial Aid, then Google and lastly the university's Hope page." Thus position $i = 1$ in the ordering is $p(1) = 1$ and $e_{p(1)p(2)} = e_{12} = 120$. The sum in the upper triangular portion of the matrix is $\text{Upper}_C(p) = 330$.

If transitions from one page to another indicate votes, then there are 330 “votes” for this ordering of page transitions and out of 720 total, the LOM = 46%. Alternately, there is a 46% agreement with the ordering $p(i) = [1,2,3,4]$. Note that an LOM of 100% would require all zeroes in the lower triangular portion of the transition matrix.

In Figure 4b, swapping the positions of Business and Google in the ordering creates $p(i) = [3,2,1,4]$. Thus position $i = 1$ in the new ordering is $p(1) = 3$ and, referring back to the c_{ij} element indexing in the original C matrix, $e_{p(1)p(2)} = e_{32} = 50$,

DOI 10.3138/infor.50.4.186
ISSN 0315-5986 | EISSN 1916-0615

$e_{p(1)p(3)} = e_{31} = 80$, $e_{p(1)p(4)} = e_{34} = 90$ and so on for the remaining elements. This increases $\text{Upper}_C(p)$ to 350 and the LOM to 49%. The optimal ordering $p(i) = [3, 4, 1, 2]$ occurs after swapping the Financial Aid and Home page, creating an $\text{Upper}_C(p) = 540$ and LOM = 75% as shown in Figure 4c. An observation supported by this example ordering is that the majority of users searched for the university in Google, went to the Home page, followed by Business pages, then to Financial Aid before exiting. However, if the site had been designed thinking that users would first check for Financial Aid, and then browse Business pages, then this data would indicate that this is not how the site is being used because the ordering shown in Figure 4d has an LOM that is 11% less than the optimal ordering.

3.2 Implementation

The steps for transforming half a million page request records into a transition matrix are shown in Figure 5. Cleansing the data requires various activities, such as removing those records associated with images displayed on a page, removing records having IPs associated with web crawlers, and removing any non-pertinent fields. The cleaned data file of records is then modified so that each page request is specified to the desired depth in the hierarchical file structure, counting the number of requests for each page and finally, eliminating those with fewer than a minimum specified number. The resulting limited log file is then sorted by IP address and time of request. This sorted file is then used to create a path file, based on the timing of page requests by an IP address. Using the paths file, transitions from one page to another can be counted and placed in a transitions matrix that is then sorted in optimal linear order by the GRASP metaheuristic.

Two key assumptions for reducing a large amount of data were: removing pages with few requests and aggregating requests to first level domains. These two assumptions reduced the number of pages in transition matrices for our data to less than 100 per day. Data records showing a single request for a page, without transitions to or from that page, were not included. For example, if a user starts a browser with its home page set to the university home page, reads any updates on that page, then quits the browser, then no paths are created and those records are not included.

SOLO results are derived from the transitions between pages, not the number of requests for a single page. In other words, the total volume of traffic for a page is not a determining factor in its ranking. For example, the Psychology department web page(s) had relatively low total traffic and was ranked as an origin. To test the effects of traffic volume on ordering, we increased Psychology traffic in the transition matrix three-fold so that it was the same magnitude as other academic departments, then re-ran SOLO, with no change occurring in the rank order, reinforcing the role distribution of traversals plays in ranking, versus the total number of page hits.

3.3 Page categorization

After calculating a linear ordering for the *n* elements of the page transition matrix, the ordering is normalized to a 1–10 scale allowing comparisons between daily traffic results with different numbers of elements. For example, the ranking of a particular page may change greatly as the traffic between various numbers of web pages changes from day to day, but on a normalized scale the ranking of a particular page may change very little. An origin is a web page ranked at the beginning of a linear ordering, hub

```
Generate_SOLO(time_period_data, depth, min_num_hits) {

   Clean_server_log ← Clean_log (time_period_data);
         // Remove extraneous data from server logs, leaving IP, date,
   time, requested URL for time period specified

   Limited_log ← Limit (Clean_server_log, depth, min_num_hits);
         // reduce URLs to the domain level & min # hits specified

   Sorted_Limited_log ← Sort (Limited_log);
         // sort by time within IP address

   Paths_log ← Create_paths (Sorted_Limited_log);
         // assign indices to URLs & create paths of indices

   Transition_matrix ← Count_transitions (Paths_log);
         // count transitions between indices

   Linear_Ordering ← GRASP (Transition_matrix);
         // rearrange matrix to maximize LOM using GRASP

end routine Generate_SOLO }
```

Figure 5. Pseudocode for Generating a Linear Ordering of Web Pages

DOI 10.3138/infor.50.4.186

pages are in the middle and destination pages towards the end of the ordering. For example, if 90 pages are ordered, then pages whose ranking is from 1 to 30 would be normalized to be from 1 to 3.3 and could be considered origins, although a strict demarcation between adjacent elements in an order should probably not be utilized.

4. SOLO RESULTS

The daily 24 hour server logs for every day from 1 January to 31 May in 2006 and 2008 were analyzed to create SOLO rankings. We present results for some representative web pages over time, then compare summarized results over time between pages. A summary traffic flow diagram for the five months is generated and compared between 2006 and 2008. The LOMs between the two years are also compared, showing that the web traffic has become more orderly.

The SOLO results over the 2006 five month time period is illustrated in Figure 6 for the home page of the university. Each daily *ranking* (position in the linear order) is categorized according to Table 1 as an Origin, Hub or Destination. The home page is ranked 69% as an Origin, 29% as a Hub, and 2% as a Destination. The three Destination rankings correspond exactly to the day before spring break, the end of spring semester, and end of summer intersession. Looking at weekends only throughout the five months, ~33% of the home page rankings are Origins, indicating the site is used differently on the weekends than on weekdays.

The university's home page is intended to be the entry point into its web site. It has many links to other important areas and many people bookmark this page. Thus most traffic is directed away from the home page and mostly high daily rankings are expected. Being ranked as a hub is also consistent since almost every web page has a link back to the university home page with its list of important links. However, for the university home page to be categorized as a destination indicates people ending their search at this page. Perhaps people are doing a final check on the "front door" before leaving, which would be consistent with the day before spring break and end of a semester which are the only days where the home page is ranked as a destination.

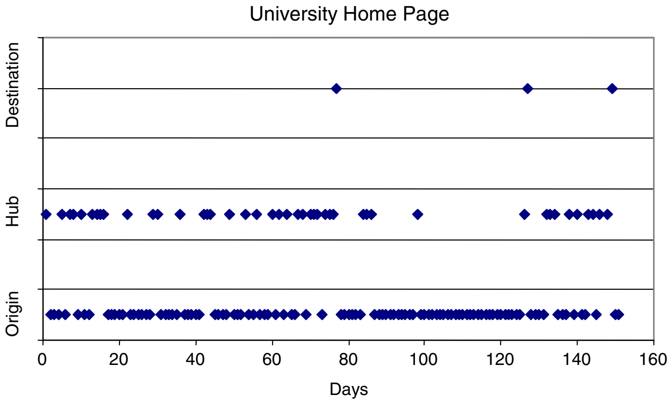

Figure 6. SOLO Categorizations for University Home in Spring 2006

TABLE 1.
Web Site categorization based on normalized SOLO rankings

Categorization	Normalized SOLO Ranking
Origin	1 < SOLO ranking ≤ 3
Hub	3 < SOLO ranking ≤ 7
Destination	7 < SOLO ranking ≤ 10

The Spring 2006 SOLO summary for pages under the Business Department directory is shown in Figure 7. The Business Department is ranked as a Hub 72% of the time and as a Destination 24%. The few times it is ranked as an Origin is during weekends. Ranking as a hub indicates people leaving from and returning to in approximate even distribution. The main business page has links to other university pages (academic units, directories, search engine) as well as useful information (programs offered, faculty, news, jobs, etc) being searched for. These characteristics are consistent with hub and destination categorizations, respectively. Business was never categorized as an origin in the 2008 data.

It is the distribution of traversals between pages that determines ranking in the linear order, not the total volume of traffic to and from a page. To test this, we choose a page whose normalized rank was 9.1. We then tripled the traffic in each element in the row and column of that page in the traversal matrix, so that its total volume was approximately the same as a page ranked at 5.2. SOLO was run using the modified transition matrix with the new rank calculated at 8.9, an insignificant change. Thus, if a web designer's goal is to change the way a page is being used, that is, to change its ranking, techniques to increase traffic volume are not appropriate.

To illustrate changes over time the normalized daily rankings for three pages were averaged for each month. Figure 8 shows that the main home page average ranking is consistently low (an origin), the Business department's average is hub-like and

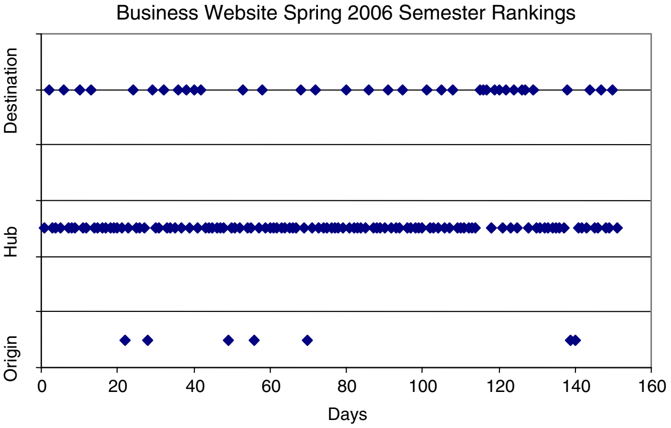

Figure 7. SOLO categorizations for Business Page in Spring 2006

ISSN 0315-5986 | EISSN 1916-0615
DOI 10.3138/infor.50.4.186

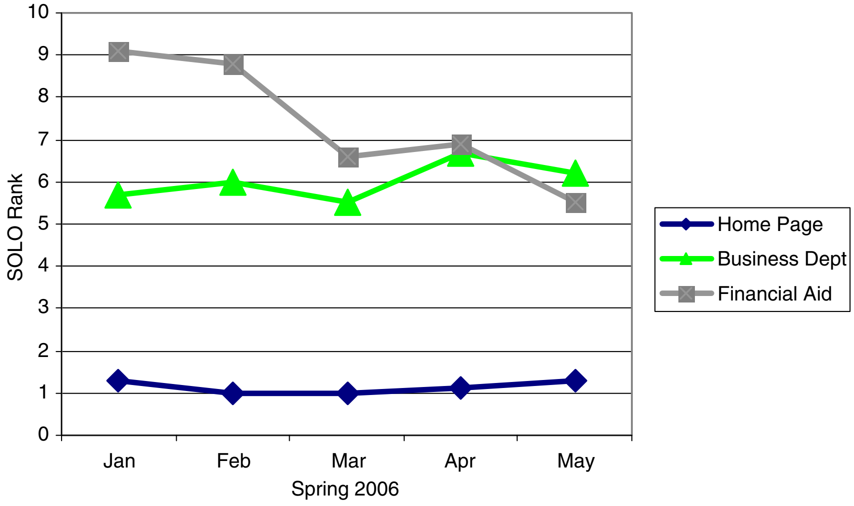

Figure 8. Average Monthly Rankings for Three Pages in Spring 2006

Financial Aid ranking changes from destination to hub as the semester progresses. Financial Aid ranked as a destination indicates that people were searching for information, then leaving the web site. Being ranked as a hub indicates traffic that returns after looking at other pages.

By taking the average of the monthly normalized rankings over the five month spring semester, a general sense of traffic flow is created. Figure 9 shows the semester's average SOLO ranking summary with sites categorized by general site content. The directed arrows in Figure 9 indicate a predominant direction of traffic, although traffic flows in all directions. Search engines and the university's main home page act as starting points to a site traversal ending with sites related to jobs, finances and social life. Sites containing lists of links, such as department listings, alphabetical listings and academic program listings are consistently ranked in the middle. These are good sites for hub activity and for helping to quickly narrow down a search. The school's private intranet can be described as a subset of the broader website with additional pages of private information. It is ranked as a hub, indicating people entering and leaving.

The summary views in Figures 9 and 10 are highly condensed and would fit on devices with small displays or limited bandwidth, such as mobile phones. They could be used to help new visitors navigate a complex, non-linear web site, helping align them to the most commonly traversed paths. It serves as a road map for users, showing them where they currently are in the web site by highlighting their current page category in the page traversal summary, as illustrated by Academic Departments in Figure 10.

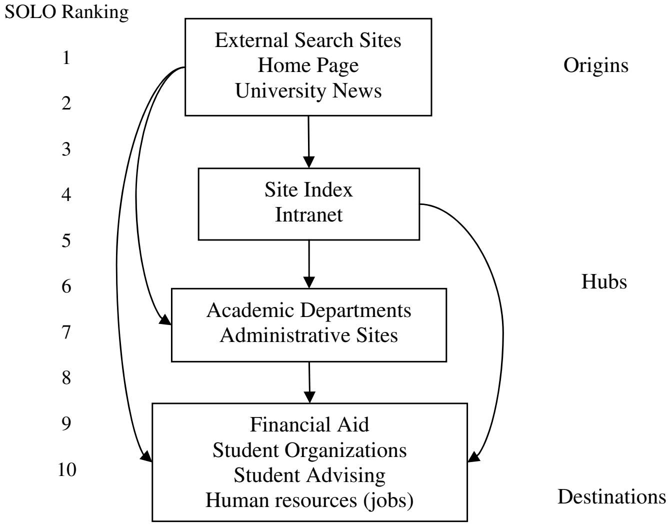

Figure 9. Average SOLO Rankings for Spring 2006

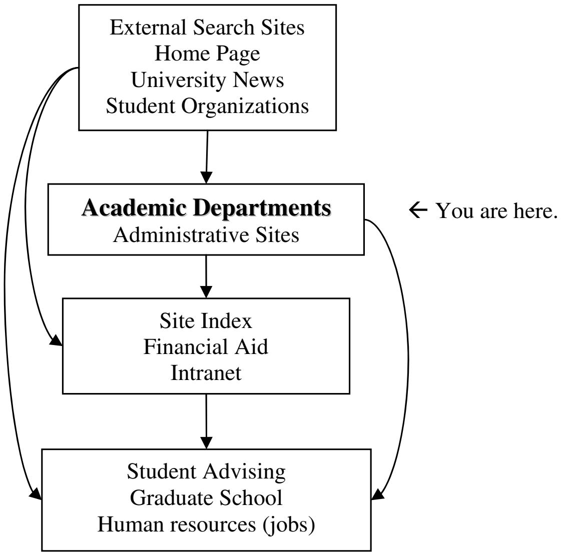

Figure 10. Average SOLO Rankings for Spring 2008 with current location highlighted

DOI 10.3138/infor.50.4.186

Comparing the page traversal summaries of 2006 and 2008, we notice consistency as well as a few changes in ranking; for example, Academic Departments and Administrative pages are still in the general hub category, but have moved closer to origin status in 2008. The password protected university intranet expanded after 2006, adding more pages accessible from it and moving it towards a destination classification. The university started graduate programs in late 2007, thus it did not show up in the spring 2006 logs. Having the graduate school categorized as a destination by SOLO in early 2008 indicates users tending to stop at, log off, or leave to another off-campus page from this directory.

From 2006 to 2008, Site Index pages moved nearer to destination classification. As people get used to the organization of a site, information is easier to find and users do not need to rely on a site index as much, therefore we would expect to see Site Index pages being used less at the beginning of a site traversal. Bookmarking frequently accessed pages, once they are found, would move them closer to origin status. After the web redesign in 2006 users (most likely students) searched for Student Organizations pages, but after being bookmarked by enough users, Student Organizations pages become an entry point origin by 2008.

4.1 Linear ordering metric results

The LOM provides a metric of traffic orderliness. The higher the LOM, the more traffic progresses in an orderly fashion towards a final goal. An example of orderly traffic is the pages visited by all users of an e-commerce site when they check out with the purchases in their shopping cart. In order from start to finish, users typically review the items in the cart, verify delivery address, select shipping method, select payment method and finally receive notification of a successful purchase with a receipt. Figure 11 compares the LOMs calculated by SOLO for each day in spring semester 2006 and 2008. We would expect that as users adjust to the reorganization of a web site, that traffic will become more orderly. Figure 11 shows that the daily LOMs for 2008 are consistently higher than in 2006. The dates were adjusted to coincide between the two years, that is, both start from the beginning day for the semester. The difference between LOMs for the two dates is .023 for the weekdays and .035 for weekends. These differences appear small, but are statistically significant at a less than 0.001% confidence interval, supporting the hypothesis that web traffic after a redesign becomes more ordered over time.

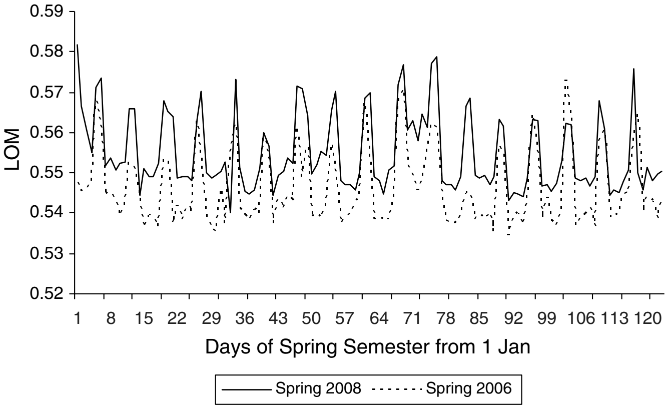

Figure 11. Average Daily LOM Increases from Spring 2006 to Spring 2008

A noticeable feature is the regular peaks and valleys. The peaks all coincide with weekends and the average weekend LOM is statistically significantly different than the weekday LOM. There is consistently less traffic on the weekends and the higher weekend LOMs could reflect a user population that is less diverse in their traversal habits.

4.2 Comparison with Markov end states

The page transition matrix can easily be converted to a stochastic matrix by dividing each element in a row by the sum of all elements in the row. The elements are then probabilities and the sum total in each row equals one. The probability of being on a given page, independent of where you started, is given by the steady state probability. The original PageRank system was basically a Markov analysis of the web crawler data used. The pages with higher steady state probabilities are ranked higher than others in the list of pages returned.

Converting the page transition matrix to a stochastic matrix then calculating the steady state probability vector, the university Home page, the entry to the university's intranet and the listing of departmental links had probabilities of .3, .16 and .05 respectively for a typical weekday. These three pages' probabilities total over 50%. Thus, over half the time users are expected to be on one of these three pages, independent of where they started. The three pages identified by Markov analysis were also identified and categorized by SOLO as an origin (Home page), destination (intranet entry page) and hub (department listings). Thus, SOLO extends our understanding of the role of these pages identified by Markov analysis.

5. CONCLUSIONS

We introduced a server-side, anonymous, non-intrusive means for collecting data and analyzing it using a new linear ordering approach. This new approach called SOLO, provides a rank ordering of the pages visited and provides summary traffic flow patterns over time. SOLO allows web designers to compare web site design goals with actual usage information. We demonstrated that some rankings change over time, while others are much more stable; for example, the ranking of the web page for student organizations changed over time while the home page is consistently ranked as an origin. We showed that orderliness of traffic is captured in the linear ordering metric. We found that the highly summarized rankings could be a useful presentation on smaller displays such as smart phones or tablets. Finally, we compared our approach to a Markov analysis and found the two approaches are compatible.

REFERENCES

Berkhin, P., Becher, J., and Randall, D. (2001), “Interactive path analysis of web site traffic”, *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’01)*, San Francisco, 414–419.

Brin, S. and Page, L. (1998), “The anatomy of a large-scale hypertextual web search engine”, *Proceedings of the 7th International World Wide Web Conference (WWW7)*, Brisbane, AU, 107–117.

Campos, V., Glover, F., Laguna, M., and Martí, R. (2001), “An experimental evaluation of a scatter search for the linear ordering problem”, *Journal of Global Optimization,* 21(4): 397–414.

Chanas, S. and Kobylanski, P. (1996), “A new heuristic algorithm solving the linear ordering problem”, *Computational Optimization and Applications,* 6: 191–205.

Cooley, R. (2003), “The use of web structure and content to identify subjectively interesting web usage patterns”, *ACM Transactions on Internet Technology,* 3(2): 93–116.

Ding, C., Zha, H., He, X., Husbands, P., and Horst, S. (2004), “Link analysis: hubs and authorities on the world wide web”, *Society for Industrial and Applied Mathematics,* 46(2): 256–268.

Huizingh, E. and Hoekstra, J. (2003), “Why do consumers like websites?”, *Journal of Targeting, Measurement and Analysis for Marketing,* 11(4): 350–361.

Laguna, M. and Martí, R. (2003), *Scatter search: Methodology and implementations in C,* Kluwer Academic Publishers, Boston.

Lewis, M., Alidaee, B., Glover, F., and Kochenberger, G. (2009), “A note on xQx as a modeling and solution framework for the linear ordering problem”, *International Journal of Operational Research,* 5(2): 152–162.

Montgomery, A., Shibo, L., Srinivasan, K., and Liechty, J. (2004), “Modeling online browsing and path analysis using clickstream data”, *Marketing Science,* 23(4): 579–595.

Reinelt, G. (1985), “The linear ordering problem: Algorithms and applications”, In H. H. H. a. R. Wille (eds), *Research and Exposition in Mathematics,* vol. 8. Heldermann Verlag, Berlin.

Spiliopoulou, M., Bamshad, M., Berendt, B., and Nakagawa, M. (2003), “A Framework for the evaluation of session reconstruction heuristics in web-usage analysis”, *INFORMS Journal on Computing,* 15(2): 171–190.

Zheng, Z., Padmanabhan, B., and Kimbrough, S. (2003), “On the existence and significance of data preprocessing biases in web-usage mining”, *INFORMS Journal on Computing,* 15(2): 148–170.

ISSN 0315-5986 | EISSN 1916-0615

DOI 10.3138/infor.50.4.186