

# Discovering interesting navigations on a web site using Sequence Alignment Method extended with an Interestingness Measure

Birgit Hay, Geert Wets and Koen Vanhoof  
Limburg University Centre, Faculty of Applied Economic Sciences,  
B-3590 Diepenbeek, Belgium  
{birgit.hay;geert.wets;koen.vanhoof}@luc.ac.be

## Abstract

In this article, a new algorithm called Sequence Alignment Method extended with an Interestingness Measure (SAM<sup>1</sup>) is illustrated for mining navigation patterns on a web site. Through log file analysis, SAM<sup>1</sup> distinguishes *interesting patterns* (i.e. unexpected, surprising patterns contradicting with the structure of the web site or direct hyperlinks between web pages) from *uninteresting patterns* (i.e. expected, known, obvious patterns resulting from the structure of the web site or direct hyperlinks between web pages) and provides information about the order of visited web pages. The algorithm is validated using real data sets of the Music Machines web site <http://machines.hyperreal.org>, home of musical electronics on the web. Empirical results show that SAM<sup>1</sup> identifies profiles of visiting behavior, which may be used for web personalization techniques and for optimizing the layout of the web site through structuring of page-links.

## 1 Introduction and Background

One of the fundamental elements in modern society is the World Wide Web (or Web), which creates a universal space of information that can be accessed by companies, government, universities, students, teachers, business people and other individuals. A web site represents a set of interconnected web pages on the Web and is developed and maintained by a person or organization. While web sites constitute a medium for communication, publicity and commerce, Web Mining studies discover and analyze useful information from the Web [Cooley *et al.*, 1997].

In general, Web Mining covers three knowledge discovery domains: Web Content Mining, Web Structure Mining and Web Usage Mining [Cooley *et al.*, 1997; Zañane, 2001]. Web Content Mining is the process of extracting knowledge from the content of documents and their descriptions. Web Structure Mining is the process of

inferring knowledge from the World Wide Web organization and links between pages in the Web. Finally, Web Usage Mining focuses on analyzing visiting information from logged data in order to extract previously unknown and interesting usage patterns [Cooley *et al.*, 1999a]. In this study, we will focus on Web Usage Mining.

Studies show that, for mining visiting behavior on web sites, different techniques are used. For example, in [Shahabi *et al.*, 2000] a tool for real-time knowledge discovery from users web page navigation called INSITE is presented. The system tracks users navigation through a web site and demonstrates real-time, scalable and adaptive clustering of navigation paths. A role-based recommendation engine allows for the web site to react to the user in real-time with customized information e.g. in target advertisement. Another technique is described in [Spiliopoulou and Faulstich, 1998] where a Web Utilization Miner (WUM) discovers interesting navigation patterns. The system consists of two modules. The Aggregation Service prepares the logged data for mining while the MINT-Processor performs the mining. Besides, MINT supports the specification of criteria of statistical, structural and textual nature. Also, an innovative aggregated storage representation for the information in the web server log is exploited by WUM.

Yet, as far as we know, no technique in Web Usage Mining studies exists that includes both a measure for distinguishing *interesting patterns* (i.e. unexpected, surprising patterns contradicting with the structure of the web site or direct hyperlinks between web pages) from uninteresting patterns (i.e. expected, known, obvious patterns resulting from the structure of the web site or direct hyperlinks between web pages) and a measure for the *order* in which pages are visited within interesting patterns. However, within Web Usage Mining research, mining navigation patterns based on interestingness and the order of visited web pages offers important

information for the purpose of supporting and increasing customer satisfaction. Such as, optimizing the layout of the web site through structuring of page-links. Therefore, we will concentrate in this study on a measure for Web Usage Mining that discovers knowledge about interesting navigations representing also structural information (or the order in which pages are visited). To this end, we will introduce Sequence Alignment Method extended with an Interestingness Measure (SAM<sup>1</sup>).

The focus of our paper is concentrated on developing and applying SAM<sup>1</sup>, rather than examining the various issues with regard to pre-processing server log data (i.e. application of specialized algorithms for sessionizing, identification of users by means of cookie registration etc.).

The article is organized as follows. First, Sequence Alignment Method (SAM) is described. Then, an Interestingness Measure, based on Baldwin's support logic [Baldwin, 1987] as well as a support logic framework for Web Usage Mining [Cooley *et al.*, 1999b], is explained. In section 4, the algorithm of SAM<sup>1</sup> is described. In section 5, SAM<sup>1</sup> is applied to a real data set storing usage behavior on the web site <http://machines.hyperreal.org>. Finally, in section 6, conclusions and topics for future research are given.

## 2 Sequence Alignment Method (SAM)

SAM is explained and illustrated in [Hay *et al.*, 2002; 2003a; 2003b]. In this section we give a short overview of the algorithm.

SAM is a distance (or similarity) measure between sequences reflecting the amount of work that needs to be done to convert one sequence into the other. The higher/lower SAM distance, the more/less effort it takes to equalize the sequences. The amount of work or effort is expressed by the following operations: insertion, deletion and reordering. Insertion and deletion operations are applied to unique elements; reordering operations are applied to common elements. *Common elements* appear in both of the compared sequences whereas *unique elements* appear in either one of them. Furthermore, *insertion* adds an element into the source (first) sequence; *deletion* removes an element from the source (first) sequence. Moreover, *reordering* changes the order of elements.

In particular, the SAM distance measure between two sequences  $S_1$  and  $S_2$  is calculated using the following formula [Joh *et al.*, 2001; Sankoff and Kruskal, 1983]:

$$d_{SAM}(S_1, S_2) = \min[(w_d D + w_i I) + \eta R] \quad (1)$$

where

- $d_{SAM}$  is the distance between two sequences  $S_1$  and  $S_2$ , based on SAM;
- $w_d$  is the weight value for the deletion operations, a positive constant not equal to 0, determined by the researcher ( $w_d > 0$ );
- $w_i$  is the weight value for the insertion operations, a positive constant not equal to 0, determined by the researcher ( $w_i > 0$ );
- D is the number of deletion operations;
- I is the number of insertion operations;
- R is the number of reordering operations;
- $\eta$  is the reordering weight, a positive constant not equal to 0, determined by the researcher ( $\eta > 0$ );

Equation (1) indicates that the score between two sequences, represented by SAM, consists of the minimum costs for deleting and inserting unique elements and reordering common elements.

To illustrate SAM, consider the following example. Sequences  $s_1$  and  $s_2$  represent sequentially ordered visited pages (each element or page within the sequence is represented by an identification number) on a web site.

Suppose:  $w_d = w_i = 1$  and  $\eta = w_d + w_i$

$s_1 \{1, 4, 7, 8\}$

$s_2 \{4, 2, 3, 1, 5\}$

First, common elements, which do not occur in the same order, are reordered. In  $s_1$ , page (or element) 4 must precede page (or element) 1 so as to be equal to the order in  $s_2$ . The result of this reordering operation is:

$s_1 \{4, 1, 7, 8\}$

$s_2 \{4, 2, 3, 1, 5\}$

Then, unique elements in  $s_1$  (i.e. pages 7 and 8) are deleted from  $s_1$  while unique elements in  $s_2$  (i.e. pages 2, 3 and 5) are inserted into  $s_1$ . Insertion in  $s_1$  is done following the order of pages in  $s_2$ . Finally, 1 reordering, 2 deletions and 3 insertions is the amount of work done to equalize  $s_1$  to  $s_2$ , resulting in a SAM distance measure  $d_{SAM}(s_1, s_2) = 7$ .

## 3 Interestingness Measure

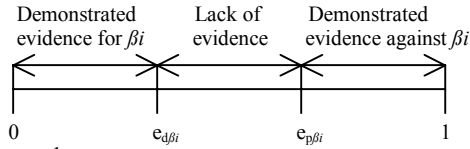
The support logic framework, used within SAM<sup>1</sup>, starts with the principles of Baldwin's support logic [Baldwin, 1987]. The framework is constructed with beliefs for Web Usage Mining [Cooley *et al.*, 1999b].

### 3.1 Baldwin's support logic

Baldwin's support logic [Baldwin, 1987] values each piece of information, also called *belief*, by the *evidence for* and *evidence against*. For each such type of evidence, two kinds of evidence definitions exist. *Demonstrated evidence* is evidence that is proven or shown by the data and known by the researcher. *Possible evidence* is evidence that is not proven by the data. The researcher

may have an idea about the existence of such evidence but it is not known for sure.

Figure 1 illustrates the conceptual frame of evidence [Baldwin, 1987; Cooley *et al.*, 1999b]. For each belief  $\beta_i$ , demonstrated evidence  $e_{d\beta_i}$  and possible evidence  $e_{p\beta_i}$  is represented by the evidence pair  $[e_{d\beta_i}, e_{p\beta_i}]$ . Furthermore, possible evidence against  $\beta_i$ , demonstrated evidence against  $\beta_i$  and lack of evidence with regard to  $\beta_i$  are represented in the framework by respectively  $(1-e_{d\beta_i})$ ,  $(1-e_{p\beta_i})$  and  $(e_{p\beta_i}-e_{d\beta_i})$ . Demonstrated as well as possible evidence must be nonnegative. Finally, summing demonstrated evidence supporting  $\beta_i$  with demonstrated evidence against  $\beta_i$  must not be greater than one.



where

$\beta_i$  = belief  $i$  with  $i = 1, 2, \dots, B$ ;

$B$  = total number of beliefs;

$e_{d\beta_i}$  = demonstrated evidence for, in support of,  $\beta_i$ ;

$e_{p\beta_i}$  = possible evidence for, in support of,  $\beta_i$ ;

$(1-e_{d\beta_i})$  = possible evidence against  $\beta_i$ ;

$(1-e_{p\beta_i})$  = demonstrated evidence against  $\beta_i$ ;

$(e_{p\beta_i}-e_{d\beta_i})$  = lack of evidence for or against  $\beta_i$ ;

$[e_{d\beta_i}, e_{p\beta_i}]$  = evidence pair of  $\beta_i$ ;

$e_{d\beta_i} \geq 0$ ;  $e_{p\beta_i} \geq 0$ ;  $e_{d\beta_i} + (1-e_{p\beta_i}) \leq 1$ ;

Figure 1. Conceptual frame of evidence.

Following Baldwin's support logic programming [Baldwin, 1987], for every belief  $\beta_i$ , evidence pairs  $[e_{d\beta_i}^1, e_{p\beta_i}^1]$  and  $[e_{d\beta_i}^2, e_{p\beta_i}^2]$ , coming from two different sources, are combined into one evidence pair  $[e_{d\beta_i}^c, e_{p\beta_i}^c]$  as follows:

$$K = 1 - e_{d\beta_i}^1(1 - e_{p\beta_i}^2) - e_{d\beta_i}^2(1 - e_{p\beta_i}^1) \quad (2)$$

$$e_{d\beta_i}^c = [e_{d\beta_i}^1 e_{d\beta_i}^2 + e_{d\beta_i}^1(e_{p\beta_i}^2 - e_{d\beta_i}^2) + e_{d\beta_i}^2(e_{p\beta_i}^1 - e_{d\beta_i}^1)] / K \quad (3)$$

$$e_{p\beta_i}^c = -\frac{[(1 - e_{p\beta_i}^1)(1 - e_{p\beta_i}^2) + (e_{p\beta_i}^1 - e_{d\beta_i}^1)(1 - e_{p\beta_i}^2) + (e_{p\beta_i}^2 - e_{d\beta_i}^2)(1 - e_{p\beta_i}^1)] / K}{1} + 1 \quad (4)$$

Interesting results are defined as either a belief with a combined evidence pair that is significantly different from (conflicting with) one of the original evidence pairs, or original evidence pairs that are significantly different from (conflicting among) each other. *Significantly different* is determined by setting a threshold value  $\tau$  for the differences between the evidence pairs. Ultimately, a belief  $\beta_i$  is interesting if:

$$\tau \leq IM_{\beta_i} \quad (5)$$

where

$$IM_{\beta_i} = \sqrt{(e_{d\beta_i}^c)^2 + (e_{p\beta_i}^c)^2} = \text{interestingness measure for } \beta_i;$$

$$e_{d\beta_i}^c = |e_{d\beta_i}^1 - e_{d\beta_i}^2|;$$

$$e_{p\beta_i}^c = |e_{p\beta_i}^1 - e_{p\beta_i}^2|;$$

### 3.2 Support logic framework for Web Usage Mining

In order to define which beliefs are interesting and which are not, we will use the two different sources of structure data and usage data, providing for each belief  $\beta_i$ , *structure evidence*  $[e_{d\beta_i}^s, e_{p\beta_i}^s]$  and *usage evidence*  $[e_{d\beta_i}^u, e_{p\beta_i}^u]$  of pages being related on a web site. A belief  $\beta_i$  is interesting if the difference between its structure and usage evidence pairs  $\geq \tau$  or if the difference between its structure (usage) and combined evidence pairs  $\geq \tau$ . We may also say that a belief  $\beta_i$  is interesting if  $IM_{\beta_i} \geq \tau$ , following equation (5).

**Calculating structure evidence.** In [Cooley *et al.*, 1999b], a method for automatically calculating structure evidence pairs for beliefs of related web pages is given. Two factors define  $e_{d\beta_i}^s$ . The *link factor* (lfactor) is a normalized measure for the number of links present among the pages of an item set. The *connectivity factor* (cfactor) is a measure for the strength of the topological connection among the pages in an item set. Structure evidence for a belief  $\beta_i$  is defined as follows:

$$e_{d\beta_i}^s = \text{lfactor} \times \text{cfactor} \quad (6)$$

where

$$\text{lfactor} = L / [P(P-1)];$$

$P$  = total number of pages in the item set;

$L$  = number of direct hyperlinks between the pages in the item set;

$\text{cfactor} = 1$  if the graphical presentation for the pages in the item set is connected, which means that minimum one direct hyperlink must exist between every pair of pages in the item set;  $\text{cfactor} = 0$  otherwise;

$e_{p\beta_i}^s$  may be set anywhere between  $e_{d\beta_i}^s$  and 1, depending on the degree of lack of evidence

**Calculating usage evidence.** In [Cooley *et al.*, 1999b], mined results from server session analyses, in the form of *frequent item sets*, representing frequently visited pages, are used to provide usage evidence for pages being related. Two measures are calculated for frequent item sets. *Support* (s) calculates the fraction of transactions that contain all of the items in the item set while *coverage* (c) measures the fraction of transactions that contain at least one of the items in the item set.

$$s = \text{count}(i_1 \square i_2 \dots \square i_p) / N \quad (8)$$

$$c = \text{count}(i_1 \square i_2 \dots \square i_p) / N \quad (9)$$

where

count (predicate) is the number of transactions containing the predicate;

$i$  is a web page in the item set;

$P$  is the total number of pages in the item set;

$N$  is the total number of transactions or server sessions;

Note that support and coverage are both highly dependent on the total number of transactions. By taking the ratio of *support-to-coverage* (SCR), this dependency is eliminated. Besides, SCR gives a single measure of the strength of a frequent item set independent of the total number of transactions in the data set. Finally,  $e_{d\beta i}^u$  is calculated as follows:

$$e_{d\beta i}^u = \text{SCR} \quad (10)$$

where

$$\text{SCR} = s / c;$$

$e_{p\beta i}^u$  may be set anywhere between  $e_{d\beta i}^u$  and 1, depending on the degree of lack of evidence

$$(11)$$

We remark that, with regard to equation (7) and (11), future research should develop an algorithm to define possible evidence for structure and usage beliefs within Web Usage Mining studies. In the following sections, no lack of evidence is tolerated [Cooley *et al.*, 1999b] and therefore  $e_{p\beta i}^s = e_{d\beta i}^s$  and  $e_{p\beta i}^u = e_{d\beta i}^u$ .

**Combining structure and usage evidence.** [Cooley *et al.*, 1999b] noticed the problem of scaling when combining structure and usage evidence into the support logic framework. Since the two sets of evidence are derived in different manners from different data sets, the scales do not necessarily match. This means that, if the number of related pages in a belief increases, the less likely it is that a corresponding frequent item set will be discovered. To deal with this, [Cooley *et al.*, 1999b] scales usage evidence based on the number of pages in the item set:

$$e_{d\beta i}^u = \text{SCR} \times \text{sfactor} \quad (12)$$

where

$$\text{sfactor} = \text{number of pages in the item set};$$

#### 4 Sequence Alignment Method extended with Interestingness (SAM<sup>I</sup>)

Beliefs of related pages that are declared interesting are integrated into the SAM algorithm. Hence, interesting visiting patterns, providing order-based information, are automatically discovered. In particular, SAM<sup>I</sup> distance between two sequences  $S_1$  and  $S_2$  is calculated using the following formula:

$$d_{\text{SAM}^I}(S_1, S_2) = \min [(w_d D^I + w_i I^I) + \eta R^I] \quad (13)$$

where

$d_{\text{SAM}^I}$  is the similarity or distance for interesting pages between two sequences  $S_1$  and  $S_2$ , based on SAM;

$w_d$  is the weight value for the deletion operations, a positive constant not equal to 0, determined by the researcher ( $w_d > 0$ );

$w_i$  is the weight value for the insertion operations, a positive constant not equal to 0, determined by the researcher ( $w_i > 0$ );

$D^I$  is the number of deletion operations for interesting pages;

$I^I$  is the number of insertion operations for interesting pages;

$R^I$  is the number of reordering operations for interesting pages;

$\eta$  is the reordering weight, a positive constant not equal to 0, determined by the researcher ( $\eta > 0$ );

Equation (13) indicates that the score between two sequences, represented by SAM<sup>I</sup>, consists of the minimum costs for deleting and inserting unique interesting elements and reordering common interesting elements.

To give a clear understanding of how SAM<sup>I</sup> works, the algorithm is illustrated with an example in table 1. The interesting beliefs of related pages, or interesting frequent item sets, discovered by the support logic framework for Web Usage Mining, are given in the first column. We remind that the order in which elements occur in frequent item sets is irrelevant. The second column presents two sequences  $s_1$  and  $s_2$  representing server sessions holding interesting and uninteresting combinations of pages. In the third column, SAM<sup>I</sup> between  $s_1$  and  $s_2$  is presented. Finally, in the last column, the original source and target sequences  $s_1$  and  $s_2$  are changed into sequences holding only interesting combinations of pages, respecting the order in which pages occur. Combinations of pages that are not interesting are filtered out of the sequences. Note that the benefit from integrating interestingness with SAM into SAM<sup>I</sup> is that server sessions are in fact ‘pre-processed’ in such a way that they only hold interesting related pages. Also, order-based information within such ‘pre-processed’ server sessions is measured by SAM<sup>I</sup>. In the end, clustering these ‘pre-processed’ server sessions based on SAM<sup>I</sup> distance measures will provide navigations that are interesting with regard to the order of occurrence of visited pages. For example, in table 1, interesting belief (2, 8) is examined by SAM<sup>I</sup> regarding order-based information within server sessions and after clustering based on SAM<sup>I</sup> distance measures, server sessions are grouped together because they hold page 8, followed by page 2 (and not the other way around).

Interesting beliefs of related pages	Source sequence: $s_1 = 2, 4, 5, 1$	$w_i = 1$ $w_d = 1$ $\eta = 2$	Source sequence based on int. bel. of related pages: $s_1 = 2, 1$
(1, 2)	Target sequence: $s_2 = 1, 3, 1, 8, 2$	$d_{SAM^I}$ ( $s_1, s_2$ ) $= 5$	Target sequence based on int. bel. of related pages: $s_2 = 1, 3, 1, 8, 2$
(7, 8)			
(2, 8)			
(1, 2, 3)			

Table 1. Sequence comparison based on SAM<sup>I</sup>.

## 5 Empirical analysis

### 5.1 Proposed approach

The empirical analysis reported in this article concerns the question whether interesting navigations providing structural information (sequential relationships or order of visited pages), embedded in web-click stream data, are well reflected by SAM<sup>I</sup>. Hence, first pair wise SAM<sup>I</sup> distances are calculated between server sessions. A *server session* represents the click stream of page views for a single user to a web site. SAM<sup>I</sup> uses its most common parameters (i.e.  $w_d = w_i = 1$ ;  $\eta = 2$ ).

Second, a distance matrix holding pair wise SAM<sup>I</sup> distance measures between server sessions is used as distance measure for clustering. For example, if 7,266 server sessions,  $s_1, s_2, s_{7,266}$  are analyzed by SAM<sup>I</sup>, distance measures are calculated between every pair of sessions i.e. between  $s_1$  and  $s_2, s_1$  and  $s_3, \dots, s_1$  and  $s_{7,266}$ , between  $s_2$  and  $s_3, s_2$  and  $s_4, \dots, s_2$  and  $s_{7,266}$ , between  $s_3$  and  $s_4, s_3$  and  $s_5, \dots, s_3$  and  $s_{7,266}, \dots$ , and finally between  $s_{7,265}$  and  $s_{7,266}$ . These distance measures are inserted into a matrix where columns and rows represent the sequences  $s_1, s_2, s_3, \dots, s_{7,266}$ . The diagonal elements of the matrix are zero because they represent the distance between equal server sessions. Finally, Ward hierarchical clustering is invoked on the matrix. In order to define the best solution for the number of clusters, a consensus among the following criteria is used. *R-squared* is used as a goodness-of-fit measure during clustering processing and equals the proportion of variation explained by the model. Furthermore, [Cooper and Milligan, 1988] have compared thirty methods for estimating the number of clusters using hierarchical clustering methods. The criteria that performed best in these simulation studies were *pseudo F statistic (PSF)* and *T-squared statistic (TST)*. Relatively large values given by the PSF indicate a stopping point. A general rule for interpreting the values of TST is to move towards joining of clusters and find values markedly larger than previous values. Finally, another method for judging the number of clusters in a data set is the *root mean squared standard deviation (RMSSTD)*, which provides a measure of homogeneity for the cluster solution. The smaller this value, the more homogeneous are the data.

Third, for each cluster, support and confidence are calculated for every combination of the order of pages within interesting frequent item sets. Here, support and confidence take into account the order of pages. *Support* is specified as the fraction of server sessions within a cluster presenting the interesting order based frequent item set. *Confidence* expresses the probability that, if a server sessions in a cluster contains all but the last page (in respective order) of the order based interesting frequent item set, the server session will also hold the last page. For each cluster, the five highest support values are used for graphically depicting interesting navigations.

### 5.2 Data

For this application, log files registering visiting behavior from 01/02/1999 till 28/02/1999 on the web site <http://machines.hyperreal.org> is analyzed. After pre-processing the data using the method described in [Hay *et al.*, 2002; 2003a], a total number of 75,855 server sessions, showing navigations through web pages with 1,159 different logged URL addresses, are identified. In a preceding step, before the actual SAM<sup>I</sup> application took place, 539 beliefs of related web pages (or frequent item sets), consisting of minimum 2 and maximum 4 related pages (or items), with a minimum support of 0.1% are defined from the usage data. For each belief, usage, structure and combined evidence pairs are calculated using equations (2) to (4) and (6) to (12). Note that no lack of evidence is tolerated in the analysis, which means that demonstrated evidence always equals possible evidence. An interestingness threshold value of  $\tau = 0.75$  in equation (5) is used to filter out interesting beliefs of related pages. By setting the value of  $\tau$  very high, related pages of the highest interest are discovered. Usually, a  $\tau$ -value of 0.5 is satisfactory [Cooley *et al.*, 1999b]. From the analysis, 91 beliefs of related pages were declared interesting.

The analysis starts with applying SAM<sup>I</sup>, taking into account the 91 interesting beliefs of related pages, to the server sessions. While calculating SAM<sup>I</sup> distance measures, the original number of server sessions is reduced to 7,266, due to the fact that SAM<sup>I</sup> selectively aligns sequences based on interesting beliefs of related pages. This means that 68,589 server sessions do not hold interesting beliefs of related pages and therefore, are not considered for further analysis.

From the data, 4 clusters are defined, following the criteria for defining the number of clusters. For each cluster, support and confidence are calculated for every combination of the order of pages within interesting beliefs of related pages. Out of a total number of 91 interesting beliefs of related pages, 278 different combinations of the order of pages are defined. For each

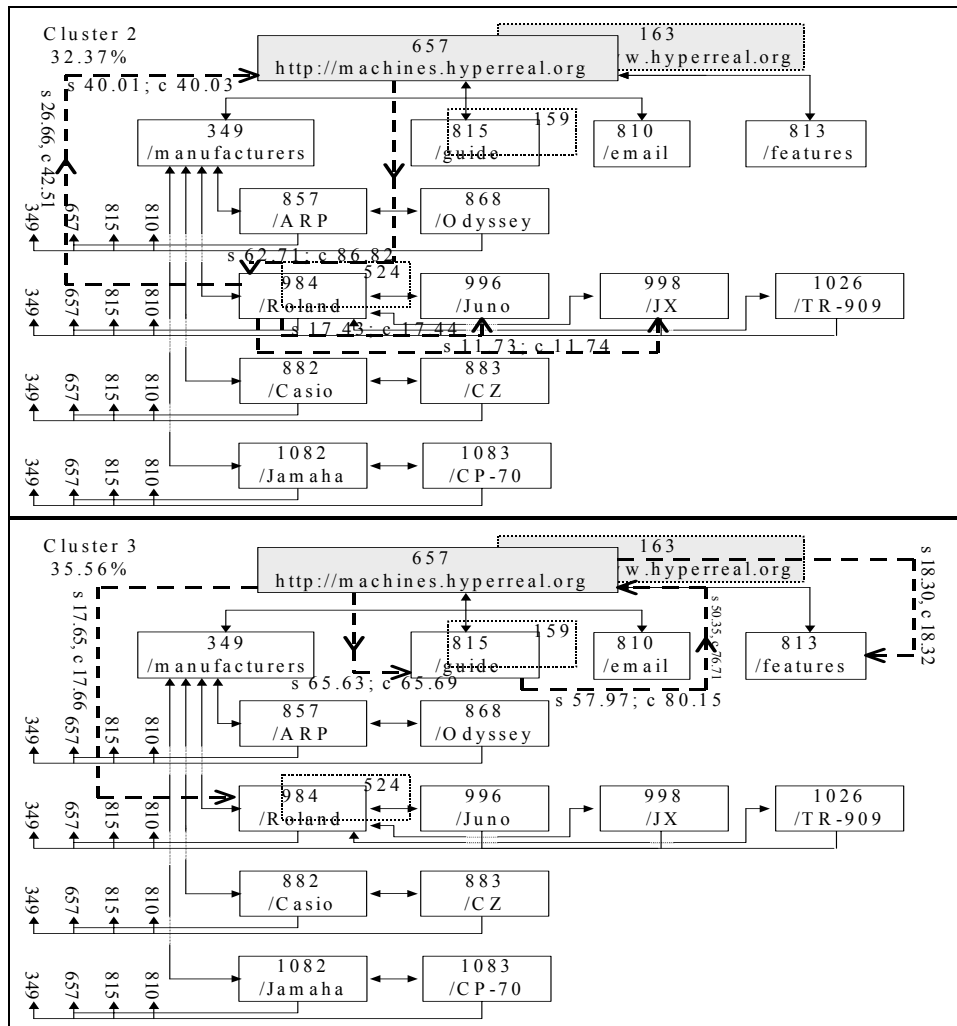


Figure 2. Interesting navigations on <http://machines.hyperreal.org>.

cluster, the 5 highest support values are used for presenting interesting navigations.

### 5.3 Results

Figure 2 presents interesting navigations providing information about the order of visited pages on the web site <http://machines.hyperreal.org>. The two largest clusters are presented, grouping together respectively 32.37% and 35.56% of the server sessions in the data set. Parts of the structure of the web site, along with direct hyperlinks between pages are graphically depicted in each cluster. For each page, the page\_id is given along with (a part of) the URL address of this particular page, which is written under the page\_id inside the rectangle. The complete URL address of each page can be read taking into account the level in the web site structure and the links. For example, page 657 constitutes the main page with URL address <http://machines.hyperreal.org>.

Going one level downwards, 4 different web pages appear. The complete URL address of page 349 is <http://machines.hyperreal.org/manufacturers>. Proceeding towards, for example, page 868, the URL address <http://machines.hyperreal.org/manufacturers/ARP/Odyssey> is given. The dashed rectangles originated from different logged URL addresses in the files. However, the content of the web page appears to be exactly the same as the one given by the solid rectangles. Further analysis revealed that the log files also stored information of people who used the URL address <http://www.hyperreal.org> and navigations from this main page on. For example, page 159 appears to be exactly the same as page 815. The only difference is that page 159 is navigated through <http://www.hyperreal.org/guide> and page 815 is navigated through <http://machines.hyperreal.org/guide>. We would like to keep this distinction in our analysis because these

From	To	Usage evidence	Structure evidence	Combined evidence	IM
349 <a href="http://machines.hyperreal.org/manufacturers">http://machines.hyperreal.org/manufacturers</a>	657 <a href="http://machines.hyperreal.org">http://machines.hyperreal.org</a>	[0.0167; 0.0167]	[1.000; 1.0000]	[1.0000; 1.0000]	1.3905
657 <a href="http://machines.hyperreal.org">http://machines.hyperreal.org</a>	813 <a href="http://machines.hyperreal.org/features">http://machines.hyperreal.org/features</a>	[0.0345; 0.0345]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3654
813 <a href="http://machines.hyperreal.org/features">http://machines.hyperreal.org/features</a>	657 <a href="http://machines.hyperreal.org">http://machines.hyperreal.org</a>	[0.0345; 0.0345]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3654
163 <a href="http://www.hyperreal.org">http://www.hyperreal.org</a>	159 <a href="http://www.hyperreal.org/guide">http://www.hyperreal.org/guide</a>	[0.0748; 0.0748]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.3084
...	...	...	...	...	...
1082 <a href="http://machines.hyperreal.org/manufacturers/Jamaha">http://machines.hyperreal.org/manufacturers/Jamaha</a>	1083 <a href="http://machines.hyperreal.org/manufacturers/JamahaCP-70">http://machines.hyperreal.org/manufacturers/JamahaCP-70</a>	[0.0995; 0.0995]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2734
984 <a href="http://machines.hyperreal.org/manufacturers/Roland">http://machines.hyperreal.org/manufacturers/Roland</a>	1026 <a href="http://machines.hyperreal.org/manufacturers/Roland/TR-909">http://machines.hyperreal.org/manufacturers/Roland/TR-909</a>	[0.1151; 0.1151]	[1.0000; 1.0000]	[1.0000; 1.0000]	1.2514
...	...	...	...	...	...

Table 2. Suggestions for reorganizing pages or deleting direct links.

navigations are considered interesting. Links between pages are drawn by *thin black solid arrows*, while interesting navigations, including order-based information, are given by the *bigger dashed arrows*. For example, from page 657, people can go to pages 349, 815, 810, 813 and from each of these pages a link points back to the home page. Also, from page 349 other pages may be visited like 857, 984, 882, 1082 as well as 657, 815 and 810. Support (s) and confidence (c) values are written next to, above or under the arrows of the interesting navigations. For example, in cluster 2, 40.01% of the server sessions visited page 984 before page 657. The confidence value indicates that, if people visit page 984, the probability that they will visit page 657 thereafter is 40.03%. Yet, permutations of the same set of pages may be presented as well. They will provide interesting information about re-visits to pages, particularly for content pages.

In order to avoid complex drawings of arrows making figure 2 unclear, some modifications are made. First, with regard to the links between pages, some arrows point towards a particular page\_id. For example, from pages 857, 984, 882, 1082 one may proceed to pages 657, 815 and 810. Likewise, from pages 868, 996, 998, 1026, 883, 1083 one may proceed to pages 349, 657, 815 and 810. Second, the dashed parts of the links indicate that there is no intersection with other links. If there were no dashed parts, the links could be misinterpreted, saying, for example, that from page 984 a link points to page 882. Third, with regard to the presentation of interesting navigations, lines showing arrows in the middle of navigations, instead of at the beginning or at the end, may appear. For example, in cluster 2, when navigating from page 657 to page 984 and from page 984 to page 996, somewhere in the middle of both navigations, an arrow is drawn. These arrows are used for interpreting order based interesting beliefs of related pages with more than 2 pages. Support and confidence values are given next to or above the

arrow of the last navigation. For example, in cluster 2, an interesting navigation appears in the following order: 657, 984, 657 with support and confidence values of 26.66% and 42.51%. Fourth, with regard to the magnitude of the structured web site with interesting related pages, for each cluster, only part of the site is given that is relevant for describing the interesting navigations.

#### 5.4 Deploying the results

Interesting navigations may provide useful information for link optimization studies. In order to develop a web site structure conform to visiting behavior of users, links between pages that are not optimally used may be deleted or pages may be moved elsewhere in the structure of the web site. Given the analysis of web usage behavior on <http://machines.hyperreal.org>, clustering server sessions based on SAM<sup>1</sup> discovered interesting navigations providing order-based information of visited pages that are used together *less* then would be expected from the structure of the web site. Some suggestions for improving the structure of the web site, based on figure 2 and on the two remaining clusters that are not shown in figure 2 due to space limitations, along with usage, structure and combined evidence, are given in table 2. Due to lack of space, we were not able to give a list of all the suggestions. In the last column, the Interestingness Measure (IM) is given. This measure may give an indication of the ‘urgency’ of reacting to the behavior of web users. The higher IM, the more urgent it is to respond to visiting behavior by optimizing the structure of the web site. For example, navigating from page 349 to 657 occurs less frequent as expected. Therefore, the direct hyperlink from page 349 to 657 may be deleted. An IM of 1.39 notifies that reaction to this behavior is most urgent.

## 6 Conclusions and Future Research

In this article, SAM is extended with an Interestingness Measure (SAM<sup>I</sup>) to discover interesting navigation patterns, providing information about the order of visited pages on a web site. Navigations are defined interesting if they are unexpected, surprising or contradicting with the structure of the web site or direct hyperlinks between web pages. The new algorithm SAM<sup>I</sup> is tested on a real data set of web usage data on <http://machines.hyperreal.org> and discovered interesting navigations that are used together *less* frequent than would be expected from the structure of the web site. This indicates that links between web pages are not optimally used and suggestions for reorganizing pages or deleting direct hyperlinks on <http://machines.hyperreal.org> may be given, along with an 'urgency' measure.

The empirical results provide 4 clusters, showing profiles of interesting navigations. Cluster 1 mainly represents navigations to and from the 'manufacturers' page. Clusters 2 and 3 represent navigations that are concentrated around the 'Roland' and 'home' page respectively. In cluster 4, interesting navigations to several pages are presented: 'guide', 'ARP', 'Odyssey', 'Casio', 'CZ', 'Jamaha' and 'CP-70'.

The clusters are well separated with regard to the order of visited pages since all of the support values of the navigations in figure 2 are below 1% for other clusters. For example, cluster 1 represents navigating from page 163 to 349. The support for this navigation in clusters 2, 3 and 4 is respectively 0.17%, 0.19% and 0.00%. One exception is made for navigating from page 657 to 815 and from page 815 back to 657. These navigations are strongly related with other interesting navigations in cluster 3 and 4. This means that server sessions in cluster 3, holding navigations 657, 815 and 815, 657, also hold navigations 657, 810; 657, 813 and 813, 657. Server sessions in cluster 4, holding navigations 657, 815 and 815, 657, also hold navigations 857, 868; 882, 883 and 1082, 1083. Finally, given the results of our study, we may conclude that interesting navigations providing structural information (sequential relationships or order of visited pages), embedded in web-clickstream data, are well reflected by SAM<sup>I</sup>.

Topics for future research should include sensitivity into IM with regard to the 'depth' of pages in the web site structure. Studies must verify whether the a-priori probability of finding related pages which are situated 'deep' in the web site structure is smaller than the probability of finding related pages which are situated at the 'top' in the web site structure. Also, an algorithm for defining possible evidence for structure and usage

beliefs in Web Usage Mining studies should be developed.

## References

- [Baldwin, 1987] J.F. Baldwin. Evidential support logic programming. *Fuzzy sets and systems*, 24(1): 1-26, 1987.
- [Cooley *et al.*, 1997] R. Cooley, B. Mobasher and J. Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web. A survey paper. In *Proc. ICTAI-97*.
- [Cooley *et al.*, 1999a] R. Cooley, B. Mobasher and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*, 1(1): 5-32, 1999.
- [Cooley *et al.*, 1999b] R. Cooley, P.-N. Tan and J. Srivastava. Discovery of interesting usage patterns from web data. Technical Report TR 99-022 University of Minnesota.
- [Cooper and Milligan, 1988] M.C. Cooper and G.W. Milligan. The effect of error on determining the number of clusters. In *Proc. Workshop on Data Analysis, Decision Support and Expert Knowledge Representation in Marketing and Related Areas of Research*, 319-328.
- [Hay *et al.*, 2002] B. Hay, G. Wets and K. Vanhoof. Web Usage Mining by means of Multidimensional Sequence Alignment Methods. In *Proc. Workshop WEBKDD-01*, 44-52.
- [Hay *et al.*, 2003a] B. Hay, G. Wets and K. Vanhoof. Mining Navigation Patterns using a Sequence Alignment Method. *Knowledge and Information Systems*, in press.
- [Hay *et al.*, 2003b] B. Hay, G. Wets and K. Vanhoof. Segmentation of visiting patterns on web sites using a Sequence Alignment Method. *Journal of Retailing and Consumer Services*, 10: 145-153, 2003.
- [Joh *et al.*, 2000] C.H. Joh, T.A. Arentze and H.J.P. Timmermans. A position-sensitive sequence alignment method illustrated for space-time activity diary data. *Environment and Planning A*, 33(2): 313-338, 2001.
- [Sankoff and Kruskal, 1983] D. Sankoff and J.B. Kruskal, editors. *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA, 1983.
- [Shahabi *et al.*, 2000] C. Shahabi, A. Faisal, F.B. Kashani and J. Faruque. INSITE: A Tool for interpreting Users? Interaction with a Web Space. In *Proc. VLDB-00*, 635-638.
- [Spiliopoulou and Faulstich, 1998] M. Spiliopoulou and L. Faulstich. WUM: a Tool for Web Utilization Analysis. In *Proc. Workshop WebDB-98*, 84-103. Extended version in *LNCS*, 1590: 84-103, 1998.
- [Zaïane, 2001] O.R. Zaïane. Conference Tutorial Notes: Web Mining: Concepts, Practices and Research. In *Proc. SDBD-2000*, 410-474.