Journal of Classification 33:298-324 (2016)
DOI: 10.1007/s00357-016-9204-8

Analysis of Web Visit Histories, Part I: Distance-Based Visualization of Sequence Rules

Roberta Siciliano

University of Naples Federico II, Italy

Antonia D'Ambrosio

University of Naples Federico II, Italy

Massimo Aria

University of Naples Federico II, Italy

Sonia Amodio

University of Naples Federico II, Italy

Abstract: This paper constitutes Part I of the contribution to the analysis of web visit histories through a new methodological framework. Firstly, web usage and web structure mining are considered as an unique mining process to detect the latent structure of the web navigation across the web sections of a single portal. We extend association rules theory to web data defining new concepts of web (patterns) association and preference matrices, as well as of (indirect and direct) sequence rules. We identify the most significant rules, according to a multiple testing procedure. In the literature, web usage patterns can be visualized in no-distance-based graphs describing the navigation behavior across web pages with sequential arrows. In the following, we introduce a geometrical visualization of sequence rules at any click of the web navigation. In particular, we provide two distance-based visualization methods for the static analysis of all data tout court and the dynamic analysis to discover the most significant web paths click by click. A real world case study is considered throughout the methodological description.

Keywords: Association rules; Sequence rules; Bonferroni inequality; Multidimensional scaling; Non-symmetric correspondence analysis.

Corresponding Author's Address: R. Siciliano, Department of Industrial Engineering, University of Naples Federico II, Naples, Italy, email: roberta@unina.it.

Published online: 25 July 2016

1. Introduction

1.1 Web Usage-Structure Mining

Web Mining can be thought of as Data Mining on a new kind of data that are available from the net in many fields of everyday life applications: communication, shopping, information, business and so on. Data can come from a single website, a group of websites or from a server (Etzioni 1996; Cooley, Mobasher, and Srivastava 1999; Kosala and Blockeel 2000; Srivastava, Cooley, Deshpande, and Pang-Ning 2000; Berry and Linoff 2002; Chakrabarti 2002). A standard format of data coming from the web does not exist: the only common characteristic of such data is that their size is typically huge (Giudici and Figini 2009; Pecoraro and Siciliano 2008). Web Mining can be distinguished into the following three main branches:

- *Web Content Mining* is the process of extracting useful information from the contents of Web documents, such as texts, graphics, or specific words like in text mining. Methods of analysis derive from Information Retrieval and Natural Language Processing.
- *Web Usage Mining* is the process of exploring large web data repositories in order to describe or predict the behavior of users, namely the usage patterns, while they are interacting with the Internet. Input data, usually derived from logfiles or tracking applications, refer to connection and visit information (time of connection, page visited, document downloaded etc.). Methods aim at identifying usage habits when visiting one or more web sites under investigation.
- *Web Structure Mining* is the process of analyzing the linkage scheme between the pages of a website, or between pages published by various websites. The traditional approach to Web Structure Mining consists of a graph in which web pages are the nodes and hyperlinks are the edges connecting two related pages. Usually, a distinction can be made between an intra-page level (when the analysis refers to a single website) and a hyperlink level (when the analysis covers two or more websites).

Web Usage and Web Structure Mining have different goals: the former aims at understanding the profile of the web user while the latter at extracting patterns from hyperlinks describing the location of the web pages. The key idea of this paper is in-between Web Usage and Web Structure Mining. More precisely, we combine these two branches into one statistical methodology that can be referred as Web Usage-Structure Mining (D'Ambrosio, Pecoraro, and Siciliano 2008). The main issue is to

Table 1. The Web Navigation Matrix (Case Study from the UCI Machine Learning Repository)

Session	First section visited	Second section visited	Third section visited	Fourth section visited	Fifth section visited	Sixth section visited	...
1	Frontpage	Sports	News	News	Weather		
2	Frontpage	Opinion	Local	Tech	Opinion	Opinion	Living
3	Weather	Travel	Tech				
4	News	News	News	Local	On-air	Frontpage	
5	BBS	Travel	Business	Travel	Living	Living	Living
6	Frontpage	Sports	Local	Sports	News	Opinion	
...	...	...	...	...	...	...	...

analyze the website as a unique entity, a unique mining process where the link between the usage and the structure is strictly related and considered together. Indeed, we consider the traditional input of Web Usage Mining process to extract information about how each page is related to another in terms of the navigation behavior, which is a typical target of Web Structure analysis. The range of the proposed methodology is a single portal. We do not consider the link structure of the web site map, but we detect the latent structure of the web navigations across the different web sections of a single portal.

1.2 The Real-World Case Study

A real world case study is considered throughout the methodological description. The dataset comes from the UCI machine learning repository. It consists of about a million navigation sessions collected in a single day on msnbc.com, an American general purpose portal and from the news related portion of msn.com. All the web-pages of this website are originally grouped into seventeen main-pages or web sections ($\{Frontpage, News, Tech, Local, Opinion, Onair, Misc, Weather, MSN-news, Health, Living, Business, MSN-sports, Sport, Summary, BBS, Travel\}$). For each navigation section, it is not possible to know how many pages are included. Notice that each navigation session is associated to a web user of the portal, which must not be confused with a specific person that can have access to the portal in different navigation sessions. Table 1 describes the structure of the data. Each row describes the clicking path of each navigation session or browsing session, registering column by column the visited pages from the entry to the website till the exit from it.

Being the number of web pages visited in each navigation session is not fixed, each row may have a different number of entries. As an example, the first navigation session goes until the fifth click on the portal, visiting respectively $\{Frontpage\}$, $\{Sport\}$, $\{News\}$, again $\{News\}$ and $\{Weather\}$; the third session goes until the third click, etc. For completeness, Figure 1 shows that about 37% of the visitors gets only one "click" on the web site, as well as more than 50% of the visitors leaves the web site after no more than two clicks. In Figure 2, it is reported the proportion of visits of each web section. Our goal is to explore the web preferences on the navigation sessions to detect the most relevant patterns in the navigation behavior.

1.3 State of the Art and Key Contributions of the Paper

The main idea of this paper is to introduce a methodological framework for web usage-structure mining by focusing the attention on the sequence of web sections that are preferred click by click until one navigator leaves the web. The aim is to discover web preferences and the most significant paths. The set of available web sections is fixed and can be related to one web portal.

The first key contribution is in terms of web preferences discovery through association and sequence rules. The presence or absence of each web section in any web navigation can be recorded to discover which pages co-occur together more often in the web navigations. For that, *association rules* (Agrawal and Srikant 1994; Zhang and Zhang 2002) can be considered. It is known that association rules are statements between two or more items in a large database of objects. In our framework, the objects are the web navigations and the items are the single web pages or suitable combinations of web pages that occur together in any web navigation (*web patterns* or *web paths* when taking into account the timing of the choices). For example, the rule $\{Frontpage, News\} \Rightarrow \{Business\}$ in web mining would indicate that if a navigator clicks on both $\{Frontpage\}$ and $\{News\}$, it is likely that also $\{Business\}$ is preferred. In the rule, the combination of web sections $\{Frontpage, News\}$ plays the role of the *Antecedent* item and $\{Business\}$ the *Consequent* item. It is worth noting that, for each web navigation, we take into account the presence or absence of each item individually selected as well as the combination of items. When the rules are temporally ordered, association rules are known as *sequence rules*, with a distinction between *indirect* and *direct* rules (Blanc and Giudici 2002): an indirect rule is when a user can choose other items between $\{Frontpage, News\}$ and $\{Business\}$, while a direct rule is when $\{Business\}$ is preferred soon after having chosen $\{Frontpage, News\}$. For that, it is necessary to record

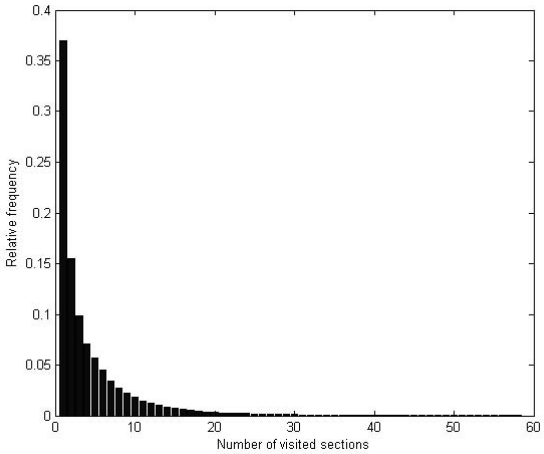

Figure 1. Bar charts of the visited sections in the case study

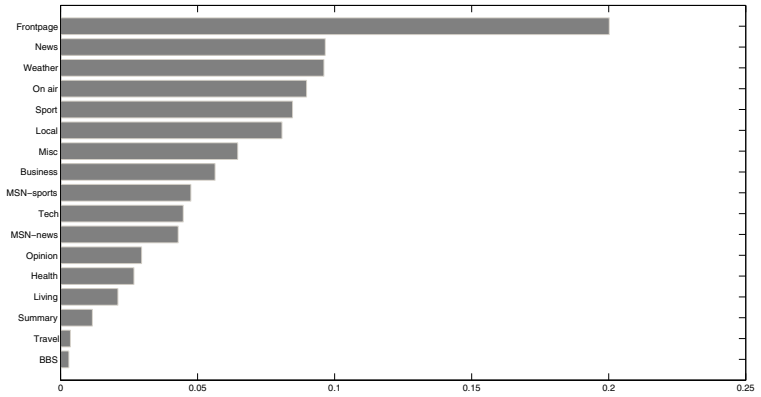

Figure 2. Proportions of visits on each web section in the case study

the web section that is visited as entry to the website at the first click, then the web section visited at the second click and so on till the exit from the website. The aim is to detect the most significant clicking paths in the web navigation. The analysis of the co-occurrences between the *Antecedent* items and the *Consequent* items is performed by considering suitable statistical measures known as support, confidence and lift. This allows the evaluation of the strength of the association (or sequence rules) in order to extract the most significant associations.

In Section 2, we introduce suitable notation and definitions of web navigation matrix and of different types of both web preferences and web

association matrices, new concepts of web patterns and indirect and direct sequence web paths. In addition, a hypothesis testing procedure is introduced to detect the most significant sequence rules.

The second key contribution is in terms of the visualization of sequence rules. So far a non-geometrical representation of sequence rules has been provided by Blanc and Giudici (2002), by drawing a line to relate two sequence rules. A more informative visualization of indirect and direct sequence rules are provided by distance-based geometrical representations by considering multidimensional scaling (Borg and Groenen 2005) and non-symmetric correspondence analysis (Lauro and Siciliano 1989; Siciliano, Mooijaart, and van der Heijden 1993) respectively. These two approaches are introduced in Section 3: the *static* analysis of all data *tout court* and the *dynamic* analysis to discover the most significant web paths click by click for web structure-usage mining.

2. Web Preferences Discovery through Association and Sequence Rules

2.1 Web Navigation Discovery: Some Basic Definitions

Let N be the total number of web navigations or browsing sessions. It is worth recalling that the total number of clicks in each navigation varies from one navigation to another. Assume that V_{max} is the highest number of clicks during any of the N navigation sessions. Let $V \leq V_{max}$ be the number of clicks to be analyzed. Typically this value is chosen to be lower than the maximum in the aim of analyzing a consistent number of web navigations.

Definition 1.1: The set $W = \{w_1, \ldots, w_j, \ldots, w_J\}$ includes J **web sections or pages** that can be visited at any navigation session and w_j is the generic web page.

Definition 1.2: Let $\{X_1, \ldots, X_v, \ldots, X_V\}$ be the set of V categorical variables describing the web preferences at any click in the web navigations, where X_v is the web preference variable at the v-th click.

Definition 1.3: The matrix $\mathbf{X} = [x_{lv}]$ $(l = 1, \ldots, N; v = 1, \ldots, V)$ is the **web navigation data matrix** of N rows and V columns, where the general entry x_{lv} denotes which web page in the set W is visited in the l-th navigation session at the v-th click, thus it can be equal to any j, for $j = 1, \ldots, J$ and $x_{lv} = 0$ if none web section is visited. By definition, the l-th navigation session includes non zero column entries till the exit from the website, while the remaining column entries are equal to zero.

2.2 Web Association and Sequence Rules: Basic Definitions

Association rules can be fruitfully considered in web mining: items are web sections and objects are web navigations. Let A be the set of antecedent items and let C be the set of consequent items. The antecedent items can be either a single web section or a combination of web sections of the set W. The consequent items are the web sections of the set W.

Definition 2.1: An **association rule** is a statement such as $w_i \Rightarrow w_j$, where the item w_i in the set A plays the role of *Antecedent* and the item w_j in the set C the role of *Consequent*.

The following statistical measures take into account the co-occurrences between the *Antecedent* item and the *Consequent* item in order to evaluate the strength of the association rules:

Definition 2.2: The **Support** of the rule $w_i \Rightarrow w_j$, denoted as $Sup_{w_i \Rightarrow w_j}$, is the proportion of navigation sessions which include simultaneously both the *Antecedent* w_i in the set A and the *Consequent* w_j in the set C. It can be also perceived as the estimate of the probability that both w_i and w_j occur in the current navigation sessions.

Definition 2.3: The **Confidence** of the rule $w_i \Rightarrow w_j$, denoted as $Conf_{w_i \Rightarrow w_j}$, is the proportion of navigation sessions which include simultaneously both the *Antecedent* w_i in the set A and the *Consequent* w_j in the set C among those that have included the *Antecedent* w_i. It can be also understood as the estimate of the conditional probability that both w_i and w_j occur given that w_i has certainly occurred in the current navigation sessions.

Definition 2.4: The **Lift** of the rule $w_i \Rightarrow w_j$, denoted as $Lift_{w_i \Rightarrow w_j}$, is the ratio between the support of the rule $w_i \Rightarrow w_j$ and its theoretical value in the case there is independence in the choice of the *Antecedent* w_i and of the *Consequent* w_j, namely the product of single supports Sup_{w_i} and Sup_{w_j}.

Support, confidence and lift measures of association rules can be deduced from the so-called web association matrix that is properly defined for any choice of the antecedent set of items. A minimum threshold of both the *Support* and the *Confidence* as well as of the *Lift* measures allow to select the most significant association rules.

For sequence rules, it is necessary to record the time of the preference because the choice of the antecedent item occurs temporally before the choice of the consequent item. A distinction is made between direct and indirect rules.

Definition 2.5: A **direct sequence rule of order v** is defined as ${w_i}^{(v-1)} \Rightarrow {w_j}^{(v)}$, where the *Antecedent* item ${w_i}^{(v-1)}$ is preferred at time $(v-1)$ and the *Consequent* item ${w_j}^{(v)}$ at time v.

Definition 2.6: An **indirect sequence rule of order v** is defined as ${w_i}^{(v')} \Rightarrow {w_j}^{(v)}$ with the time $v' < v$ so that the item ${w_i}^{(v')}$ is temporally *Antecedent* to the *Consequent* item ${w_j}^{(v)}$.

For this reason, the set of indirect rules includes the set of direct rules (when $v' = v-1$) and all other sequence rules between a couple of items where the antecedent item can be chosen at any other click $v' < v-1$.

Support, confidence and lift measures for sequence rules can be easily derived taking into account the time of the occurrences of the antecedent and consequent items. Respect to common association rules problems, in our case the total number of object to consider is huge. Indeed, the current navigations for sequence rules are those navigators that are still active at the v-th click. For sequence rules a hypothesis testing procedure is introduced to select the most significant ones.

2.3 Web Preference and Association Matrices

The information contained in the the web navigation matrix can be summarized by recording the presence or absence of each web section in the navigations to derive the web preferences as well as the co-occurrences between any couple of web sections. For that we define the web preference matrix and the web association matrix to derive the support, confidence and lift measures of the association rules.

Definition 3.1: Let the matrix $\mathbf{Z} = [z_{lj}]$ $(l = 1, \ldots, N; j = 1, \ldots, J)$ be the **web preference matrix** of N rows and J columns, where the general entry is $z_{lj} = 1$ if the web page w_j of the set W has been visited at least once in the l-th navigation session, and $z_{lj} = 0$ otherwise.

Definition 3.2: Let the matrix $\mathbf{S} = \mathbf{Z}'\mathbf{Z} = [s_{ij}]$ be the J square **web association matrix**, where the general entry s_{ij} $(i, j = 1, \ldots, J)$ describes the co-occurrence of the i-th row antecedent web section of the set W with the j-th column consequent web section of the set W in the N web navigations.

The entries of the web association matrix allows to derive the support measures of the association rules $w_i \Rightarrow w_j$ between any couple of row antecedent web section and column consequent web section, such as $Sup_{w_i \Rightarrow w_j} = s_{ij}/N$ for $i, j = 1, \ldots, J$.

The diagonal terms of the matrix $\mathbf{S}$ of general term s_{ii} (for $i = 1, \ldots, J$) denote the occurrences of each antecedent web section in the current web navigations. The support measures of the row antecedent web sections can be derived as $Sup_{w_i} = s_{ii}/N$ for $i = 1, \ldots, J$.

The confidence measures of the j-th consequent web section given that the i-th row antecedent web section has been preferred can be obtained as $Conf_{w_i \Rightarrow w_j} = s_{ij}/s_{ii}$ $(j = 1, \ldots, J)$. It describes the proportion of times each consequent web section occurs given that the i-th row antecedent web section has been preferred.

By definition, the lift measures of any association rule can be derived as $Lift_{w_i \Rightarrow w_j} = N s_{ij}/s_{ii}s_{jj}$ for $i, j = 1, \ldots, J$.

2.4 Web Patterns Preference and Association Matrices

The association rules can be also defined for an antecedent item formed by a combination of web sections from the set W and the consequent being any web section in the set W. For that, we introduce the concept of web pattern to define the web patterns preference matrix and the web patterns association matrix.

Definition 4.1: The **web pattern of order h** for any $1 < h < J$ is any combination (without replacement) of h web sections from the J elements in the set W that might occur together. Let $W_{(h)}$ be the set of all possible web patterns of order h with cardinality equal to $I_{(h)} = \binom{J}{h}$.

Definition 4.2: The matrix $\mathbf{Z}_{(h)} = [z_{li_{(h)}}]$ is the **web patterns preference matrix of order h** $(1 < h < J)$ of N rows and $I_{(h)}$ columns, where the general entry is $z_{li_{(h)}} = 1$ if the $i_{(h)}$-th web pattern of order h has occurred in the l-th active navigation session, and $z_{li_{(h)}} = 0$ otherwise.

Definition 4.3: The matrix $\mathbf{S}_{(h)} = \mathbf{Z}'_{(h)}\mathbf{Z} = [s_{i_{(h)}j}]$ is the **b patterns association matrix of order h**, where the general entry $[s_{i_{(h)}j}]$ describes the co-occurrence of the $i_{(h)}$-th row web pattern of order h of the set $W_{(h)}$ with the j-th column web section of the set W in the N web navigation sessions. The entries of this matrix allows to derive, up to a scaling factor, the support measures of the association rules where the antecedent item is a web pattern of order h of the set $W_{(h)}$ and the consequent item is any web section in the set W. Confidence and lift measures can be derived straightforwardly.

It is worth noting that these definitions are given for completeness to emphasize a formal description of how to derive statistical measures of sup-

port, confidence and lift for any type of association rules with combination of web sections, not necessarily bound to the longitudinal nature of the data matrix.

2.5 Web Longitudinal Preferences and Association Matrices

The web navigation data matrix can be also summarized in terms of longitudinal web preference matrix describing the presence or absence of each web section preference sequentially click by click, so that the co-occurrences and association rules can be analyzed specifically for any combination of clicks.

Definition 5.1: The matrix $\tilde{\mathbf{Z}} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_v, \ldots, \mathbf{Z}_V]$ of N rows and $J \times V$ columns is the **longitudinal web preference matrix**, which is obtained by juxtaposing the V sub-matrices of the type $\mathbf{Z}_v$ of N rows and J columns that describe in disjoint coding the web preference variable X_v at the v-th click, for $v = 1, \ldots, V$. In other words, the general entry of the sub-matrix $\mathbf{Z}_v$ is $z_{lj(v)} = 1$ if the web page w_j of the set W has been visited in the l-th navigation session at the v-th click, and $z_{lj(v)} = 0$ otherwise. It is worth noting that at any click v, $N_{(v)}$ navigators are still active and they can choose any web section in the set W, with $N_{(1)} = N$. For any click $v > 1$ the matrix $\mathbf{Z}_v$ will include $(N - N_{(v)})$ zero rows for all left out navigators.

The co-occurrences between any couple of web sections in the set W click by click can be recorded by the longitudinal web association matrix.

Definition 5.2: The matrix $\tilde{\mathbf{S}} = \tilde{\mathbf{Z}}'\tilde{\mathbf{Z}}$ is the $J \times V$ square **longitudinal web association matrix**. This matrix is formed by V square block sub-matrices of dimension J of the type $\mathbf{S}_{v',v} = \mathbf{Z}'_{v'}\mathbf{Z}_v = [s_{i(v')j(v)}]$ $(i, j = 1, \ldots, J)$ where the general entry $s_{i(v')j(v)}$ $(i, j = 1, \ldots, J; v', v = 1, \ldots, V)$ describes the co-occurrence of the i-th row antecedent web section preferred at the v'-th click with the j-th column consequent web section preferred at the v-th click in the web navigations. It can be shown that $\sum_i \sum_j s_{i(v')j(v)} = N_{(v)}$.

By definition, the matrix $\tilde{\mathbf{S}}$ is symmetric; the diagonal blocks matrices for $v' = v$ are diagonal matrices which general entries divided by $N_{(v)}$ provide the support measures of each web section at the v-th click. The superior triangular blocks of $\tilde{\mathbf{S}}$ allows to derive the support measures for all *indirect sequence rules between single web sections*, i.e., ${w_i}^{(v')} \Rightarrow {w_j}^{(v)}$, being

$$Sup_{{w_i}^{(v')} \Rightarrow {w_j}^{(v)}} = s_{i(v')j(v)}/N_{(v)}, \tag{1}$$

$$Sup_{w_j{}^{(v)}} = s_{+j(v)}/N_{(v)}, \tag{2}$$

where $s_{+j(v)} = \sum_i s_{ij(v)}$ holds.

Considering just those matrices with $v' = v - 1$ we can derive the support measures for the *direct sequence rules between single web sections click by click*, i.e., $w_i{}^{(v-1)} \Rightarrow w_j{}^{(v)}$. As an example, at click $v = 3$ the entries of the block association matrix $\mathbf{S}_{2,3}$ divided by $N_{(3)}$ provide the support measures of direct rules between the row antecedent web section preferred at second click and the column consequent web section preferred at the third click.

Confidence and lift measures can be derived in a straightforward way:

$$Conf_{w_i{}^{(v')} \Rightarrow w_j{}^{(v)}} = s_{i(v')j(v)}/s_{i(v')+}\,, \tag{3}$$

where $s_{i(v')+} = \sum_j s_{i(v')j(v)}$, and

$$Lift_{w_i{}^{(v')} \Rightarrow w_j{}^{(v)}} = N_{(v)} s_{i(v')j(v)}/s_{i(v')+} s_{+j(v)}. \tag{4}$$

2.6 Web Paths Preferences and Association Matrices

Direct and indirect sequence rules can also be defined for the antecedent formed by a combination of web sections temporally chosen before the v-th click whereas the consequent is any web section in the set W preferred at the v-th click.

For the direct rules it is necessary to record the web sections preferred click by click yielding to the so-called direct web path.

Definition 6.1: The **direct web path of order (v-1)** $(v = 2, \ldots, V)$ is any combination (with replacement) of $(v - 1)$ web pages from the J elements in the set W being preferred before the v-th click. Let $\tilde{W}_{(v-1)}$ be the set of all possible direct web paths of order $(v - 1)$ with cardinality equal to $\tilde{I}_{(v-1)} = J^{(v-1)}$.

The corresponding web paths preference and association matrices can be defined as follows:

Definition 6.2: The matrix $\tilde{\mathbf{Z}}_{(v-1)} = [\tilde{z}_{li_{(v-1)}}]$ is the **direct web paths preference matrix of order (v-1)** of $N_{(v-1)}$ rows and $\tilde{I}_{(v-1)}$ columns, where the general entry is $\tilde{z}_{li_{(v-1)}} = 1$ if the i-th web path of order $(v - 1)$ has occurred in the l-th active navigation session, and $\tilde{z}_{li_{(v-1)}} = 0$ otherwise. By definition, $\tilde{\mathbf{Z}}_{(1)} = \mathbf{Z}_1$.

Definition 6.3: The matrix $\tilde{\mathbf{S}}_v = \tilde{\mathbf{Z}}'_{(v-1)}\mathbf{Z}_v = [\tilde{s}_{i_{(v-1)}j(v)}]$ of $\tilde{I}_{(v-1)}$ rows and J columns is the v**-th direct web paths association matrix** where the general entry $\tilde{s}_{i_{(v-1)}j(v)}$ describes the co-occurrence of the i-th row antecedent direct web path of order $(v-1)$ with the j-th column consequent web page at the v-th click in the active $N_{(v)}$ web navigation sessions. The entries divided by $N_{(v)}$ allow to evaluate the support measures of direct sequence rules where the antecedent items are web paths. Confidence and lift measures can be derived in a straightforward way.

The indirect web path needs to be defined in order to consider any combination of web sections preferred before the v-th click.

Definition 6.4: The **indirect web path of order (v-1)** $(v = 2, \ldots, V)$ is any combination (with replacement) of $v' < v$ web pages from the J elements in the set W that have been preferred before the v-th click. Let $\tilde{\tilde{W}}_{(v-1)}$ be the set of all possible indirect web paths of order $(v-1)$ with cardinality equal to $\tilde{\tilde{I}}_{(v-1)} = \sum_{v'=1}^{v-1} J^{v'} \binom{v-1}{v'}$.

The corresponding web paths preference and association matrices can be defined as follows:

Definition 6.5: The matrix $\tilde{\tilde{\mathbf{Z}}}_{(v-1)} = [\tilde{\tilde{z}}_{li_{(v-1)}}]$ is the **indirect web paths preference matrix of order (v-1)** of $N_{(v-1)}$ rows and $\tilde{\tilde{I}}_{(v-1)}$ columns, where the general entry is $\tilde{\tilde{z}}_{li_{(v-1)}} = 1$ if the i-th indirect web web path of order $(v-1)$ has occurred in the l-th active navigation session, and $\tilde{\tilde{z}}_{li_{(v-1)}} = 0$ otherwise.

Definition 6.6: The matrix $\tilde{\tilde{\mathbf{S}}}_v = \tilde{\tilde{\mathbf{Z}}}'_{(v-1)}\mathbf{Z}_v = [\tilde{\tilde{s}}_{i_{(v-1)}j(v)}]$ of $\tilde{\tilde{I}}_{(v-1)}$ rows and J columns is the v-th **indirect web paths association matrix** where the general entry $[\tilde{\tilde{s}}_{i_{(v-1)}j(v)}]$ describes the co-occurrence of the i-th row antecedent indirect web path of order $(v-1)$ with the j-th column consequent web page at the v-th click in the active $N_{(v)}$ web navigation sessions. The entries divided by $N_{(v)}$ allow to evaluate the support measures of indirect sequence rules where the antecedent items are indirect web paths. Confidence and lift measures can be derived in a straightforward way.

2.7 Significance Assessment of Sequence Rules

In the proposed methodology, the sequence rules ensuring a minimum threshold value of the support measure are considered for selecting the most

significant measures according to a hypothesis testing procedure. This can be performed on the lift measure that takes into account the deviation of the support measure from the theoretical value when the Antecedent and the Consequent items are independent. On this purpose, the binomial test can be used to verify the statistical significance of the lift measure (Hämäläinen 2010). The choice of the significance value α is harder in data mining with respect to the classical testing procedure due to the large number of patterns to be tested and to the multiple testing problem. As a result, an exhaustive search over all possible patterns needs to be performed. Typically, the significance value is corrected by considering a multiple comparisons test with the Bonferroni adjustment (Dunn 1961; Shaffer 1995; Abdi 2007a), where the desired significance level α is divided by the number of tests q. The q value can be interpreted as the number of possible rules (and possible tests) that can be generated by the discovery pattern process. When testing the sequence direct rules of order $(v-1)$ it holds $q = \tilde{I}_{(v-1)}$; when testing the indirect sequence rules of order $(v-1)$ it holds $q = \tilde{\tilde{I}}_{(v-1)}$.

Under the null hypothesis, it is assumed that *Lift* ≤ 1. For this reason we consider significant the rules for which *Lift* > 1 at a significance level equal to α/q accordingly to Bonferroni adjustment, where usually $\alpha = 0.05$. Let $N_{(v)}$ be the sample size corresponding to the number of active navigation sessions at any click v, the *Lift* can be expressed also as:

$$\frac{N_{(v)} \times Sup_{w_i \Rightarrow w_j}}{N_{(v)} \times Sup_{w_i} \times Sup_{w_j}}.$$

According to the Binomial test, the probability of success is given by $\pi_0 = Sup_{w_i} \times Sup_{w_j}$ under the null hypothesis with mean expectation $\mu_0 = N_{(v)} \times \pi_0$ and variation

$$\sigma^2 = N_{(v)} \times \left(Sup_{w_i} \times Sup_{w_j}\right)\left(1 - \left(Sup_{w_i} \times Sup_{w_j}\right)\right).$$

The sampling estimate is given by $\overline{X} = N_{(v)} \times Sup_{w_i \Rightarrow w_j}$. In our experiment, as $\pi_o \to 0$ and $N_{(v)} \to \infty$, we consider the Poisson approximation to perform the test:

$$z_0 = \frac{\left(\overline{X} - \mu_0 + 0.5\right)}{\sqrt{\mu_0}},$$

for which $Pr\left(Z \leq z_0\right) \approx \Phi\left(z_0\right) - \frac{1}{6\sqrt{\mu_0}}\left(z_0^2 - 1\right)\phi\left(z_0\right)$.

Table 2 and Table 3 show respectively the significant indirect and direct sequence rules by assuming $V = 40$. We set a minimum support value equal to 1% and we select those rules that have a lift measure significantly higher than 1.

For sake of brevity, these results are interpreted in Section 3, in combination with the visualization of indirect and direct sequence rules as provided by the proposed methods.

Table 2. Support, confidence and lift measures of the significant indirect sequence rules (threshold $\alpha = 0.05\%$).
Fp=Frontpage; Ne=News; Te=Tech; Lo=Local ; Op=Opinion; Oa=On air; Mi=Misc; We=Weather; Mn=MSN-News; He=Health; Li=Living; Bu=Business; Ms=MSN-Sports; Sp=Sport; Su=Summary; Bb=BBS; Tr=Travel

Antecedent	Consequent	Support	Confidence	Lift
Fp	Fp	0.106	0.608	2.753
Fp	Ne	0.054	0.507	1.554
Fp	Te	0.017	0.559	1.558
Fp	Li	0.014	0.623	1.402
Ne	Ne	0.043	0.405	4.254
Te	Te	0.013	0.450	15.101
Lo	Lo	0.053	0.543	6.793
Op	Op	0.028	0.647	19.354
Oa	Oa	0.028	0.434	6.229
Oa	Mi	0.016	0.170	2.289
Mi	Fp	0.017	0.092	1.255
Mi	Lo	0.011	0.112	1.252
Mi	Oa	0.012	0.187	1.991
Mi	Mi	0.043	0.449	4.893
We	We	0.090	0.813	7.412
Mn	Lo	0.012	0.125	1.575
Mn	Mi	0.011	0.124	1.561
Mn	Mn	0.015	0.539	13.948
He	He	0.015	0.461	14.304
Li	Fp	0.012	0.070	1.855
Bu	Bu	0.025	0.493	9.856
Ms	Ms	0.021	0.672	17.604
Ms	Sp	0.012	0.125	2.734
Sp	Sp	0.050	0.516	6.948
Fp & Fp	Fp	0.070	0.414	3.487
Fp & Ne	Fp	0.013	0.077	1.230
Fp & Ne	Ne	0.021	0.194	3.639
Fp & Li	Fp	0.011	0.063	2.009
Ne & Ne	Ne	0.022	0.209	5.024
Te & Te	Te	0.011	0.346	21.988
Lo & Lo	Lo	0.038	0.376	7.925
Op & Op	Op	0.022	0.503	19.558
Oa & Oa	Oa	0.015	0.239	7.961
Oa & Oa	Mi	0.010	0.101	2.551
Oa & Mi	Oa	0.010	0.162	5.497
Oa & Mi	Mi	0.013	0.127	5.585
Mi & Mi	Mi	0.021	0.224	5.096
We & We	We	0.079	0.708	7.538
Mn & Mn	Mn	0.011	0.406	19.593
He & He	He	0.012	0.352	20.106
Bu & Bu	Bu	0.018	0.354	13.899
Ms & Ms	Ms	0.015	0.511	18.487
Ms & Sp	Sp	0.010	0.107	8.063
Sp & Sp	Sp	0.031	0.311	7.486
Fp & Fp & Fp	Fp	0.052	0.307	4.172
Fp & Ne & Ne	Ne	0.015	0.138	4.560
Ne & Ne & Ne	Ne	0.019	0.180	5.888
Lo & Lo & Lo	Lo	0.034	0.333	8.476
Op & Op & Op	Op	0.020	0.457	19.933
Oa & Oa & Oa	Oa	0.013	0.194	9.920
Oa & Mi & Mi	Mi	0.011	0.108	6.127
Mi & Mi & Mi	Mi	0.016	0.164	6.321
We & We & We	We	0.074	0.668	7.788
He & He & He	He	0.011	0.335	23.763
Bu & Bu & Bu	Bu	0.016	0.311	16.137
Ms & Ms & Ms	Ms	0.013	0.423	21.287
Sp & Sp & Sp	Sp	0.026	0.257	8.120
Fp & Fp & Fp & Fp	Fp	0.043	0.251	4.550
Fp & Ne & Ne & Ne	Ne	0.012	0.113	5.542
Ne & Ne & Ne & Ne	Ne	0.018	0.171	6.087
Lo & Lo & Lo & Lo	Lo	0.031	0.302	8.774
Op & Op & Op & Op	Op	0.019	0.419	20.210
Oa & Oa & Oa & Oa	Oa	0.011	0.174	10.890
Mi & Mi & Mi & Mi	Mi	0.013	0.137	7.370
We & We & We & We	We	0.070	0.637	7.917
Bu & Bu & Bu & Bu	Bu	0.015	0.285	17.182
Ms & Ms & Ms & Ms	Ms	0.011	0.382	22.818
Sp & Sp & Sp & Sp	Sp	0.021	0.212	8.528

Table 3. Support, confidence and lift measures of the direct sequence rules (the significant rules are identified by an asterisk for a threshold $\alpha = 0.05\%$).
Fp=Frontpage; Ne=News; Te=Tech; Lo=Local ; Op=Opinion; Oa=On air; Mi=Misc; We=Weather; Mn=MSN-News; He=Health; Li=Living; Bu=Business; Ms=MSN-Sports; Sp=Sport; Su=Summary; Bb=BBS; Tr=Travel

Antecedent	Consequent	Support	Confidence	Lift	Sig
Fp	Fp	0.171	0.384	1.8	*
Fp & Fp	Fp	0.085	0.500	2.5	*
Fp & Fp	Ne	0.019	0.108	1.0	
Fp & Fp & Fp	Fp	0.060	0.704	3.9	*
Fp & Fp & Fp & Fp	Fp	0.047	0.779	4.4	*
Fp	Ne	0.059	0.132	1.7	*
Fp & Ne	Ne	0.031	0.534	4.9	*
Fp & Ne	Fp	0.017	0.287	1.1	
Fp & Ne & Ne	Ne	0.021	0.658	5.8	*
Fp & Ne & Ne & Ne	Ne	0.015	0.713	6.5	*
Fp	Li	0.031	0.070	1.9	*
Fp & Li	Fp	0.019	0.602	3.1	*
Fp	Sp	0.029	0.066	1.0	
Fp	Oa	0.027	0.062	0.9	
Fp	Bu	0.026	0.058	1.0	
Fp	Mi	0.025	0.057	0.8	
Fp	He	0.018	0.041	1.1	
Fp	Te	0.018	0.039	1.1	
Fp	Lo	0.016	0.036	0.5	
Bu	Bu	0.023	0.581	10.2	*
Bu & Bu	Bu	0.019	0.753	6.8	*
Bu & Bu & Bu	Bu	0.016	0.878	17.2	*
Bu & Bu & Bu & Bu	Bu	0.015	0.926	18.5	*
Lo	Lo	0.041	0.753	10.1	*
Lo & Lo	Lo	0.036	0.883	10.9	*
Lo & Lo & Lo	Lo	0.033	0.914	10.6	*
Lo & Lo & Lo & Lo	Lo	0.031	0.944	10.3	*
Ms	Ms	0.034	0.568	16.3	*
Ms & Ms	Ms	0.026	0.765	18.7	*
Ms & Ms & Ms	Ms	0.021	0.778	21.1	*
Ms & Ms & Ms & Ms	Ms	0.016	0.803	23.7	*
Mn	Mn	0.031	0.387	8.9	*
Mn & Mn	Mn	0.019	0.604	14.5	*
Mn & Mn & Mn	Mn	0.012	0.643	20.1	*
Mn	Mi	0.012	0.150	2.0	*
Mn	Fp	0.011	0.133	0.6	
Ne	Ne	0.025	0.608	6.1	*
Ne & Ne	Ne	0.019	0.753	6.8	*
Ne & Ne & Ne	Ne	0.014	0.775	6.8	*
Ne & Ne & Ne & Ne	Ne	0.012	0.808	7.3	*
Oa	Oa	0.037	0.393	5.3	*
Oa & Oa	Oa	0.020	0.536	8.1	*
Oa & Oa & Oa	Oa	0.014	0.689	11.3	*
Oa & Oa & Oa & Oa	Oa	0.010	0.774	11.8	*
Oa	Mi	0.021	0.212	2.9	*
Oa & Mi	Mi	0.017	0.880	9.6	*
Oa & Mi & Mi	Mi	0.014	0.803	8.6	*
Oa & Mi & Mi & Mi	Oa	0.010	0.773	11.1	*
Op	Op	0.018	0.874	33.8	*
Op & Op	Op	0.010	0.966	29.6	*
Sp	Sp	0.025	0.781	11.5	*
Sp & Sp	Sp	0.022	0.877	11.2	*
Sp & Sp & Sp	Sp	0.020	0.912	10.3	*
Sp & Sp & Sp & Sp	Sp	0.019	0.929	10.1	*
We	We	0.089	0.889	8.2	*
We & We	We	0.083	0.938	8.4	*
We & We & We	We	0.080	0.962	8.5	*
We & We & We & We	We	0.078	0.965	8.4	*

3. Distance-Based Visualization of Sequence Rules

3.1 Static Analysis: Visualization of Indirect Sequence Rules

Static analysis aims to detect the web navigation behavior of all N navigators across the web sections, eventually taking into account the longitudinal preferences click by click. The main issue is *to visualize indirect sequence rules between single web sections*.

A distance-based visualization of indirect sequence rules is provided by considering Non-Metric MultiDimensional Scaling (MDS) (Borg and Groenen 2005). In this way, it is possible to understand the similarities and the differences among browsing sections of the web navigation behavior. MDS detects similarities or dissimilarities within a set of items (in our case, the web sections) through a p-dimensional factorial plot (with typically $p = 2$). The data entry of MDS models is a proximity (either similarity or dissimilarity) matrix that typically comes from a "direct" measurement. As an alternative, we consider web data that are collected indirectly by an automatic system (tracking) used by the regular server activity (D'Ambrosio and Pecoraro 2011), thus data comes from "implicit" opinions, made manifest by the clicks of the visitors in the navigation sessions. We define this concept of similarity as *implicit behavior concordance* because it allows to implicitly deduce users behavior with respect to the main-pages. In our case, the similarity is the co-occurrence of any couple of web sections in the set W such as expressed by the web association matrix $\mathbf{S}$ of general term s_{ij} $(i, j = 1, \ldots, J)$. It can be also interesting to analyze the web longitudinal association matrix $\tilde{\mathbf{S}}$ through the general block matrices $\mathbf{S}_{v',v}$ $(v', v = 1, \ldots, V)$ with general entry $s_{i(v')j(v)}$ $(i, j = 1, \ldots, J; v', v = 1, \ldots, V)$ to detect the co-occurrences click by click. Co-occurrence data can be seen as direct proximities (Borg and Groenen 2005), but in this case some of the web pages are much more popular than others. In such a case, direct analyses of co-occurrences can yield uninteresting solutions. For such a reason, we collect dissimilarities from co-occurrence data by using the gravity model (Tobler and Wineburg 1971; Borg, Groenen, and Mair 2013). It is worth noting that the diagonal entries of the web association matrix $\mathbf{S}$ as well as of the v-th diagonal block matrix $\mathbf{S}_{v',v}$ for $v' = v$ are, up to a scaling factor, the support measures of each web section, thus measuring the popularity of each web section as well as of each web section at each click respectively.

In the analysis of the web association matrix $\mathbf{S}$, the aim is to visualize the degree of similarity of the different web sections, without taking into account the clicking time. Let δ_{ij} be the elements of the dissimilarity matrix

Table 4. Goodness and badness of fit measures for the Non-metric MDS in the case study

STRESS and Fit Measures (web association matrix)	
Normalized raw STRESS	0.042
STRESS-I	0.205
D.A.F.	0.958
Tucker's Coefficient of Congruence	0.978
STRESS and Fit Measures (longitudinal web association matrix)	
Normalized raw STRESS	0.078
STRESS-I	0.280
D.A.F.	0.922
Tucker's Coefficient of Congruence	0.960

$\mathbf{D}$ for MDS. Under the gravity model in non-metric MDS $\delta_{ij} = \left(\frac{s_{ii}s_{jj}}{s_{ij}}\right)^{\frac{1}{2}}$. Specifically, $1/s_{ij}$ transforms the co-occurrence (which is a similarity) into a dissimilarity measure (i.e., usually MDS works with dissimilarities), the multiplication by $s_{ii}s_{jj}$ standardizes the measure for the popularities of the respective web sections, and the square root follows from the physical law which justifies the gravity model, and it is not relevant for non-metric MDS.

Analogously, we can perform a longitudinal analysis click by click by considering the matrix $\mathbf{S}_{v',v}$. It is worth nothing that for co-occurrences equal to zero (namely zero entries in either the matrix $\mathbf{S}$ or in any block matrix $\mathbf{S}_{v',v}$ for $v' \neq v$), the dissimilarity measure cannot be defined. Thus, zero entries are treated as missing proximities.

We consider a non-metric Multidimensional Scaling by using the PROXSCAL algorithm (Heiser 1988; Commandeur and Heiser 1993) and treating the co-occurrences themselves as weights. Note that PROXSCAL uses the SMACOF algorithm (De Leeew 1977) as optimization engine, which minimizes the normalized raw STRESS. STRESS is a loss function that allows to find the best solution to represent the proximities among items in a subspace. To avoid a premature stop of the algorithm due to the presence of local minima, we used a configuration with 500 random start points. For more details, we refer to both Borg and Groenen (2005) and Cox and Cox (2001).

Figures 3 and 4 show the two-dimensional MDS representations of the web association matrix and of the longitudinal web association matrix in the case study fixing $V = 40$. Goodness and badness of fit of both the analyses are reported in Table 4.

The normalized raw STRESS is the proportion of variation of the dissimilarities not accounted for by the distances. The D.A.F. index is a coefficient of determination indicating the fitted proportion. Tucker's congru-

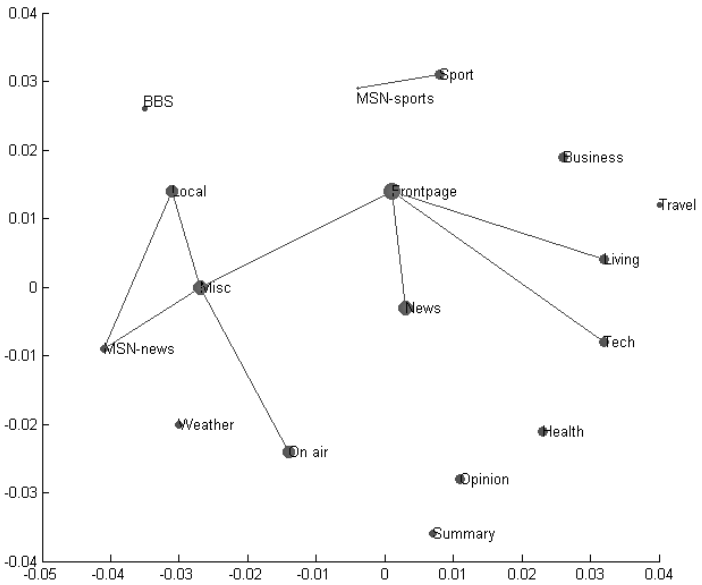

Figure 3. Non-metric MDS visualization of the web association matrix in the case study

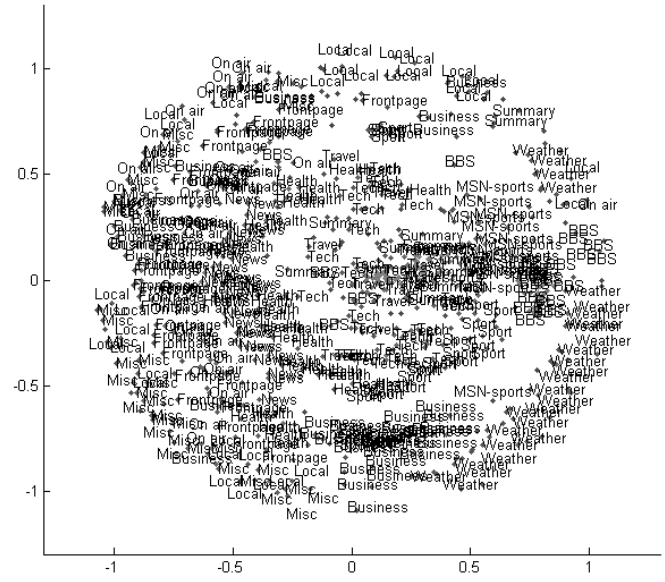

Figure 4. Non-metric MDS visualization of the longitudinal web association matrix (V=40 clicks) in the case study

ence coefficient (Abdi 2007b) is a measure of similarity between two square symmetric matrices, such as the distances matrix and the dissimilarities matrix. The square of Tucker's congruence coefficient can be understood as explained variance measure (Borg and Groenen 2005; De Leeuw 1977). It is worth noting that for both MDS solutions both the STRESS and the fit measures are quite good: the STRESS measures are close to zero as well as the fit measures are close to one.

Points in Figure 3 are proportional to the support measure of each web section and lines denote the significant indirect rules between two single web sections. It is remarkable that only few significant indirect rules involve two distinct web sections. In other words the most of the web navigators keeps standing on the same web page and rarely surfers tend to move among different sections.

The surfers' behaviour can be better emphasized by looking at Figure 4 that shows the MDS solution for the co-occurrences of the web-sections for all 40 clicks. This confirms the results of Table 2. Notice that surfers always tend to visit the same sections. Specifically, they keep standing on the same web page after the third click. In Figure 4, the sections for which there are significant indirect rules ($\{Frontpage, News\}$; $\{Local, Misc\}$; $\{Local, On-air\}$; $\{Onair, Frontpage\}$; and so on) are mostly positioned on the left side, while on the right one there are two major "clusters" ($\{Weather\}$, $\{Business\}$) and the interchanges between $\{Sport\}$ and $\{MSN-sports\}$). The sections that are scattered throughout the geometric space (such as $\{Travel\}$, $\{BBS\}$, $\{Summary\}$) are not characterized by significant indirect rules.

3.2 Dynamic Analysis: Visualization of Direct Sequence Rules

Dynamic analysis aims to detect the web preferences of the active navigators click by click, thus taking into account the temporally choice of the web section when passing from one click to another. The main issue is *to visualize direct sequence rules considering the direct sequence web path preferences*.

The visualization of direct sequence rules is performed through a factorial method for the analysis of the dependence in two-way cross-classification such as Non-Symmetric Correspondence Analysis (NSCA) (Lauro and Siciliano 1989, 2000) of the **v-th direct web paths association matrix $\tilde{\mathbf{S}}_v$** ($v = 2, \ldots, V$) of general entry $\tilde{s}_{i(v-1)j(v)}$. Specifically, the column variable can be considered as the response variable with J categories in the set W at click v and the row variable the (compound) predictor with $\tilde{I}_{(v-1)}$ categories in the set of all direct sequence paths at click $(v-1)$. For sake of brevity, in the following we denote the response variable as Y

and the predictor as X, as well as the generic co-occurrence of the i-th row antecedent web path at click $(v-1)$ and the j-th column consequent web section at click v as $\tilde{s}_{i(v-1)j(v)} = \tilde{s}_{ij}$.

NSCA analyzes the centered matrix of row profiles $\tilde{s}_{j|i} = \tilde{s}_{ij}/\tilde{s}_{i+}$ through a generalized singular value decomposition, which can be written in scalar form as:

$$\tilde{s}_{j|i} - \tilde{s}_{+j} = \sum_{k=1}^{K} \lambda_k r_{ik} c_{jk}, \tag{5}$$

where $K \leq min(I, J) - 1$ and singular values λ_k are such that $\lambda_1 \geq \ldots \lambda_k \ldots \lambda_K > 0$, with row and column scores satisfying centering and orthonormality conditions. The aim is to approximate the centered matrix by a reduced-rank decomposition with a number of factors K^* lower that K, i.e., $K^* = 2$ to obtain a bi-dimensional factorial representation. The coordinates of the row categories (i.e., the web paths playing the role of Antecendent) are defined by $\lambda_k r_{ik}$ and the coordinates of the column categories (i.e., the web sections playing the role of Consequent) are defined by c_{jk}. In this way, it can be ensured that the graphical display in the reduced space is a *biplot*. Justification for NSCA lies in the Goodman and Kruskal's predictability index $\tau_{Y|X}$ that is defined by the ratio between the explained heterogeneity due to the predictor's modalities and the total heterogeneity of the response variable (i.e., the Gini diversity index):

$$\tau_{Y|X} = \frac{\sum_i \sum_j \left(\tilde{s}_{j|i} - \tilde{s}_{+j}\right)^2}{1 - \sum_j \tilde{s}_{+j}^2}, \tag{6}$$

which varies between 0 and 1. It can be interpreted as the relative increase in correct predictions of the response variable when the knowledge about the predictor's categories is used. NSCA decomposes the numerator of the predictability index $\tau_{Y|X}$ along principal axes and also over the row and the column categories:

$$(1 - \sum_j \tilde{s}_{+j}^2)\tau_{Y|X} = \sum_{k=1}^{K} {\lambda_k}^2 = \sum_i \tilde{s}_{i+} \sum_k \left(\lambda_k r_{ik}\right)^2 = \sum_j \sum_k \left(\lambda_k c_{jk}\right)^2. \tag{7}$$

It is worth noting that the components $\sum_k \left(\lambda_k \tilde{s}_{j|i}\right)^2$ and $\sum_k \left(\lambda_k c_{jk}\right)^2$ are respectively equal to the squared distances of the row categories and of the column categories to the origin. It is possible to detect which row categories contribute more to the explained heterogeneity of the response variable as well as which column categories are best predicted by the predictor. Furthermore, the row coordinates $\lambda_k r_{ik}$ are related to the column coordinates c_{jk} by the following transition formula:

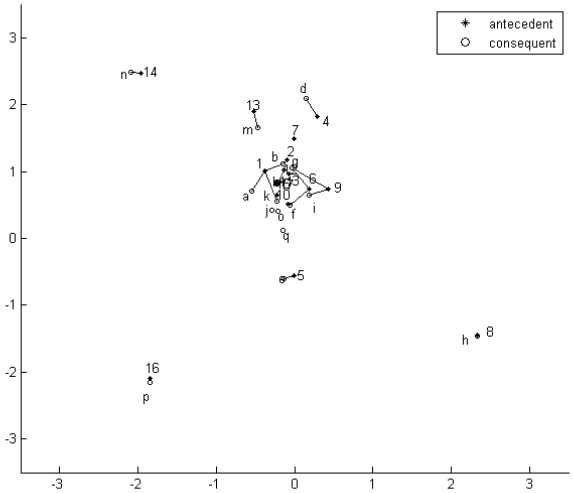

Figure 5. NSCA visualization of the Direct Sequence Rules of order V=2.
Cumulated inertia: 0.521. Goodman and Kruskall τ: 0.314 Gini diversity index: 0.908.
Antecedent items: 1=Frontpage; 2=News; 3=Tech; 4=Local; 5=Opinion; 6=On air; 7=Misc; 8=Weather; 9=MSN-news; 10=Health; 11=Living; 12=Business; 13=MSN-sports; 14=Sport; 15=Summary; 16=BBS; 17=Travel
Consequent items: a=Frontpage; b=News; c=Tech; d=Local; e=Opinion; f=On air; g=Misc; h=Weather; i=MSN-news; j=Health; k=Living; l=Business; m=MSN-sports; n=Sport; o=Summary; p=BBS; q=Travel

$$\sum_j (\tilde{s}_{ji} - \tilde{s}_{+j}) c_{jk} = \sum_j \tilde{s}_{j|i} c_{jk} - \sum_j \tilde{s}_{+j} c_{jk} = \lambda_k r_{ik} \,. \tag{8}$$

This shows that the row coordinates are, apart from a constant term, the weighted average of the column coordinates (the so-called *barycentric property*). In the case of perfect prediction, for each column point j there exists at least one row point i with the same coordinates, that is $\lambda_k r_{ik} = c_{jk}$. Thus, the predictor category i and the response category j are projected into the same point. Under independence $\lambda_k r_{ik} = 0$ for all rows, since $\tilde{s}_{j|i} = \tilde{s}_{+j}$ for all i and j. In practice, the predictive power of the row categories is somewhere between the extremes of independence and perfect prediction. Row points far from the origin have high predictive power on the response variable. A row category i with coordinates $\lambda_k r_{ik}$ has high influence in predicting a particular column category j when the column coordinates c_{jk} are high and their signs agree with the signs of the respective row coordinates. In a factorial representation this means that the row point i is close to the column point j and these points are relatively far from the origin. Figure 5 provides NSCA visualization of the direct sequence rules when passing

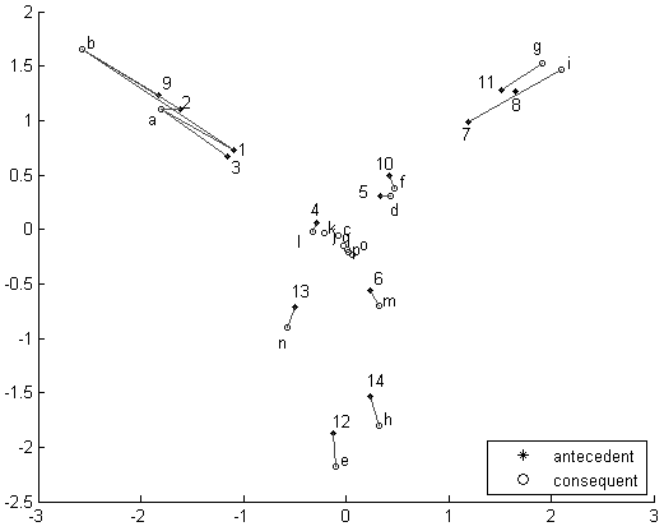

Figure 6. NSCA visualization of the Direct Sequence Rules of order V=3.
Cumulated inertia: 0.501. Goodman and Kruskall τ: 0.504 Gini diversity index: 0.892.
Antecedent items: 1=Frontpage & Frontpage; 2=Frontpage & News; 3=Frontpage & Living; 4=Business & Business; 5=Local & Local; 6=MSN-sports & MSN-sports; 7=MSN-news & MSN-news; 8=MSN-news & Misc; 9=News & News; 10=On air & On air; 11=On air & Misc; 12=Opinion & Opinion; 13=Sport & Sport; 14=Weather & Weather
Consequent items: a=Frontpage; b=News; c=Tech; d=Local; e=Opinion; f=On air; g=Misc; h=Weather; i=MSN-news; j=Health; k=Living; l=Business; m=MSN-sports; n=Sport; o=Summary; p=BBS; q=Travel

from the first click to the second click. Figure 6 provides NSCA visualization of the direct sequence rules when passing from the second click to the third click. Figure 7 provides NSCA visualization of the direct sequence rules when passing from the third click to the fourth click. In each plot lines connecting an antecedent web path with the consequent web section denote which direct sequence rules are significant according to Table 3.

From Figure 5 we can note that surfers going at the first click on $\{Weather\}$, $\{BBS\}$ and $\{Sport\}$ keep standing on the same page at the second click. Indeed, those points are very close in the plot and far from the origin. Significant rules at the second click are those starting from $\{Frontpage\}$.

In Figure 6, we can note that when passing from the second to the third click, there is a group of surfers keeping standing on the same page and another group that reach the same consequent web page from different significant web paths. As an example: $\{Frontpage, Frontpage\} \Rightarrow \{Frontpage\}$, $\{Frontpage, Living\} \Rightarrow \{Frontpage\}$; $\{Frontpage, News\} \Rightarrow \{Frontpage\}$. The convex hull of the consequent web sections

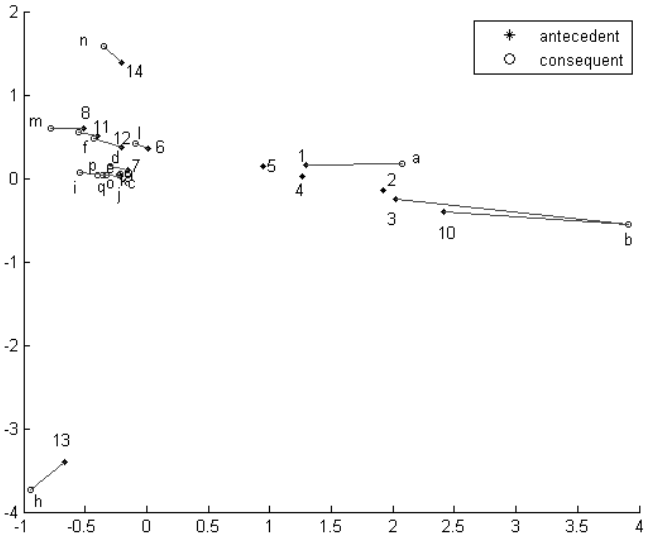

Figure 7. NSCA visualization of the Direct Sequence Rules of order V=4.
Cumulated inertia: 0.10. Goodman and Kruskall τ: 0.615 Gini diversity index: 0.878.
Antecedent items: 1=Frontpage & Frontpage & Frontpage; 2=Frontpage & Frontpage & News; 3=Frontpage & News & News; 4=Frontpage & News & Frontpage; 5=Frontpage & Living & Frontpage; 6=Business & Business & Business; 7=Local & Local & Local; 8=MSN-sports & MSN-sports & MSN-sports; 9=MSN-news & MSN-news & MSN-news; 10=News & News & News; 11=On air & Misc & Misc; 12=On air & On air & On air; 13=Weather & Weather & Weather; 14=Sport & Sport & Sport
Consequent items: a=Frontpage; b=News; c=Tech; d=Local; e=Opinion; f=On air; g=Misc; h=Weather; i=MSN-news; j=Health; k=Living; l=Business; m=MSN-sports; n=Sport; o=Summary; p=BBS; q=Travel

contains all antecedent web paths because of the barycentric property of NSCA.

Passing to the fourth click, in Figure 7 it is remarkable noting that we detect the most robust direct sequence paths rules, namely $\{Frontpage, News, News\} \Rightarrow \{News\}$, $\{News, News, News\} \Rightarrow \{News\}$.

4. Concluding Remarks and Generalization

This paper has provided a new methodological framework for web usage-structure mining. Throughout the methodological description a real world data set from the UCI Machine Learning Repository has been considered. The key ideas were to define suitable data matrices:

- the web navigation data matrix to describe the clicking path of each navigation session;

- the web preference data matrix to describe the visited web pages in the navigation sessions;
- the web association matrix to describe the co-occurrence of any couple of web pages in the navigation sessions;
- the web paths preferences and association matrices for longitudinal analysis describing the choice click by click.

As a result, sequence rules based on direct and indirect web paths preferences have been defined. A multiple testing procedure has been considered to assess the most significant ones.

Distance-based visualization of indirect and direct sequence rules have been proposed considering MDS and NSCA respectively. So far, the literature provided site-maps in which the web sections were described by circle or nodes and hyperlinks were the edges connecting two related pages (or sections), discarding any possible representation in a geometric space. As an alternative, the proposed distance-based visualization methods provide a non-random configuration of the web sections (or the web pages) in the geometric space. Specifically, static analysis has been introduced to detect the web navigation behavior of all navigators across the web sections, eventually taking into account the longitudinal preferences click by click. The main result is to visualize in a factorial space indirect sequence rules between single web sections. Dynamic analysis has been introduced to detect the web preferences of the active navigators click by click, thus taking into account the temporally choice of the web section when passing from one click to another. The main result is to visualize in a factorial space direct sequence rules by considering the direct sequence web path preferences.

In this paper, we focalize the attention on web mining, for that we have chosen a well known data set to describe the proposed methodology. In particular, we have underlined the innovative elements compared to classical approaches for the identification of significant sequence rules of browsing data and for their visualization in a geometric space. Our proposal can be used in several different real contexts of applications in both world wide web and more generic socio-economic contexts.

Thanks to the spread on a large scale of electronic devices (smartphone, tablet, ultrabook, etc.) always connected to the internet, the e-commerce has an annual growth of about 20%. The firms working in this business store users data about surfing both in the single sessions and the timing of their logins. Fixing a specific period (last session, a week, a month, etc.), it is possible to identify a sequence of web pages visited by a user, the clicks on the same pages, the items visualized and eventually bought. The analysis of these data using our proposal could support the firms' management by giving information about which browsing patterns are more prob-

ably linked to the purchase of a specific item, about which patterns are the most interesting, about which items and pages are more correlated, etc.

Another very interesting application is the analysis of data collected by supermarkets: instead of web navigators we can think of transactions at a supermarket. The concept of web pattern can be very useful to understand which combination of antecedent items is more related to a consequent target item. With respect to the concept of web path, the web pattern is not related to the timing of the choice. The development of technologies of RFID and intelligent baskets allows to store transaction data that are more complex rather than the classic matrix of presence/absence of items as in Market Basket Analysis (Al-Safadi 2010). In this way, supermarkets can collect the ordering selection sequence of products in the shopping cart and the pattern followed by the customer during the purchasing session (Hurjui, Graur, and Turcu 2008). The data collected by these technologies represent the ideal matrix for the usage of our methodology, taking into account that a longitudinal analysis can also be performed.

References

ABDI, H. (2007a), "Bonferroni and Šidák Corrections for Multiple Comparisons", in *Encyclopedia of Measurement and Statistics*, ed. N.J. Salkind, Thousand Oaks, CA: Sage, pp. 104–108.

ABDI, H. (2007b), "RV Coefficient and Congruence Coefficient", in *Encyclopedia of Measurement and Statistics*, ed. N.J. Salkind, Thousand Oaks, CA: Sage, pp. 850–856.

AL-SAFADI, L.A.E. (2010), "A Dual-Mode Intelligent Shopping Assistant", *Advances in Information Sciences and Service Sciences, 2(4)*, 43–54.

AGRAWAL, R., and SRIKANT R. (1994), "Fast Algorithms for Mining Association Rules", in *Proceedings of the 20th International Conference on Very Large Data Bases*, Santiago, Chile, pp. 487–499.

BERRY, M.J.A., and LINOFF, G.S. (2002), *Mining the Web: Transforming Customer Data*, New York: John Wiley and Sons.

BLANC, E., and GIUDICI, P. (2002), "Sequence Rules for Web Clickstream Analysis", *Advances in Data Mining, Lecture Notes in Computer Science, 2394/-1*, 1–14.

BORG, I., and GROENEN, P.J.F. (2005), *Modern Multidimensional Scaling*, New York: Springer-Verlag.

BORG, I., GROENEN, P.J.F., and MAIR, P. (2013), *Applied Multidimensional Scaling*, Heidelberg: Springer.

CHAKRABARTI, S. (2002), *Mining the Web*, San Francisco CA: Morgan Kaufmann.

COMMANDEUR, J.J.F., and HEISER, W.J. (1993), "Mathematical Derivations in the Proximity Scaling (PROXSCAL) of Symmetric Data Matrices", Technical Report No. RR-93-03, Leiden University, The Netherlands, Department of Data Theory.

COX, A., and COX, T.F. (2001), *Multidimensional Scaling*, London: Chapman and Hall.

COOLEY, R., MOBASHER, B., and SRIVASTAVA, J. (1999), "Data Preparation for Mining World Wide Web Browsing Patterns", *Knowledge and Information Systems, 1*, 5–32.

D'AMBROSIO, A., and PECORARO, M. (2011), "Multidimensional Scaling as Visualization Tool of Web Sequence Rules", in *Classification and Multivariate Analysis for Complex Data Structures, Studies in Classification, Data Analysis, and Knowledge Organization*, eds. B. Fichet et al., Berlin, Heidelbert: Springer-Verlag, pp. 307-314.

D'AMBROSIO, A., PECORARO, M., SICILIANO, R. (2008) "Web Preferences Visualization through Multidimensional Scaling and Trees", *DATAVIZ VI International Conference on Statistical Graphics: Data and Information Visualization in Today's Multimedia Society*, Bremen, Jacobs University, June 25-28, 2008 (Organizers: Lars Linsen and Adi Wilhelm).

DE LEEUW, J. (1977), "Application of Convex Analysis to Multidimensional Scaling", in *Recent Developments in Statistics*, eds. J.R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, Amsterdam: North Holland Publishing, pp. 133–145.

DUNN, O.J. (1961), "Multiple Comparisons Among Means", *Journal of the American Statistical Association, 56,* 52–64.

ETZIONI, O. (1996), "The World Wide Web: Quagmire or Gold Mine", in *Communications of the ACM, 39(11)*, 65–68.

FREUND, Y., and SCHAPIRE, R.E. (1997), "A Decision-Theoretic Generalization of Online Learning and an Application to Boosting", *Journal of Computer and System Sciences, 55(1)*, 119–139.

GIUDICI, P., and FIGINI, S. (2009), *Applied Data Mining for Business and Industry*, New York: Wiley.

HÄMÄLÄINEN, W. (2010), "StatApriori: An Efficient Algorithm for Searching Statistically Significant Association Rules" *Knowledge and Information Systems, 23(3)*, 373–399.

HASTIE T., TIBSHIRANI R., and FRIEDMAN J. (2009), *The Elements of Statistical Learning* (2nd ed.), Springer-Verlag.

HEISER, W.J. (1988), "PROXSCAL, Multidimensional Scaling of Proximities", in *International Meeting on the Analysis of Multiway Data Matrices, Software Guide*, eds. A. Di Ciaccio and G. Bove, Rome: C.N.R., pp. 77–81.

HURJUI, C., GRAUR, A., and TURCU, C.O. (2008), "Monitoring the Shopping Activities from the Supermarkets Based on the Intelligent Basket by Using RFID Technology", *Electronics and Electrical Engineering, 3(83),* 7–10.

LAURO, N.C., and SICILIANO, R. (1989), "Exploratory Methods and Modelling for Contingency Tables: An Integrated Approach", *Statistica Applicata: Italian Journal of Applied Statistics, 1,* 5–32.

LAURO, N.C., and SICILIANO, R. (2000), "Analyse non symmetrique des correspondances pour des tables de contingences", in *L'Analyse des Correspondances et les techniques connexes, partie III*, eds. J. Moreau, P.A. Doudin, and P. Cazes, Berlin, Heidelberg: Springer Verlag, pp. 183–210.

KOSALA, R., and BLOCKEEL, H. (2000), "Web Mining Research: A Survey", *ACM SIGKDD Explorations, 2*, 1–15.

PECORARO, M., and SICILIANO, R. (2008), "Statistical Methods for User Profiling in Web Usage Mining", in *Handbook of Research on Text and Web Mining Technologies*, eds. M. Song, and Y.B. Wu, Hershey PA: Idea Group Inc., pp. 359–368.

SHAFFER, J. (1995), "Multiple Hypothesis Testing", *Annual Review of Psychology, 46,* 561–584.

SICILIANO, R., MOOIJAART, A., and VAN DER HEIJDEN, P.G.M. (1993),"A Probabilistic Model for Nonsymmetric Correspondence Analysis and Prediction in Contingency Tables", *Journal of Italian Statistical Society, 2(1)*, 85–106.

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., and TANS, P.-N. (2000), "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations, 1*, 12–23.

TOBLER, W., and WINEBURG, S. (1971), "A Cappadocian Speculation", *Nature, 231*, 39–41.

ZHANG, C., and ZHANG, S. (2002), *Association Rule Mining: Models and Algorithms*, Heidelberg: Springer-Verlag.