MCCSIS 2010
IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS
Freiburg, GERMANY 28-31 JULY

Proceedings of the IADIS International Conference Intelligent Systems and Agents 2010 and European Conference Data Mining 2010

EDITED BY
Antonio Palma dos Reis
and Ajith P. Abraham
IADIS INTERNATIONAL CONFERENCE

INTELLIGENT SYSTEMS AND AGENTS 2010

and

IADIS EUROPEAN CONFERENCE ON DATA MINING 2010

part of the

IADIS MULTI CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS 2010
PROJECTION BASED SAMPLING FOR MORE EFFICIENT HIGH UTILITY ITEMSET MINING
Alva Erwin, Raj P. Gopalan and N.R. Achuthan

AS3: A FRAMEWORK FOR AUTOMATIC SLEEP STAGE SCORING
Tim Schlüter, Timm Kifelts and Stefan Conrad

SHORT PAPERS

PREVENTIVE DIAGNOSTICS FOR CARDIOVASCULAR DISEASES BASED ON PROBABILISTIC METHODS AND DESCRIPTION LOGIC
Björn-Helge Busch, Ralph Welge, Christian Weiβ and Markus Bette

MUTINY: A MULTI-TIME INTERVAL PATTERN DISCOVERY APPROACH TO PRESERVE THE TEMPORAL INFORMATION IN BETWEEN
Alessio Bertone, Tim Lammarsch, Thomas Turic, Wolfgang Aigner, Silvia Miksch and Johannes Gärtner

A THREE-TIERED WEB BASED EXPLORATION AND REPORTING TOOL FOR DATA MINING
Ahmet Selman Bozkir and Ebru Akcapinar Sezer

STUDY OF WEBSITE STRUCTURE PATTERNS FROM THE PERSPECTIVE OF SOCIAL NETWORK ANALYSIS
B. Palacios, M. R. Martinez Torres, S. L. Toral and F. Barrero

OPTIMAL PERIMETER OF THE NEIGHBOURHOOD IN FUZZY DECISION TREE
Youdthachai Lertworaprapchaya and Yingjie Yang

MODULAR DATA HIDING FOR DIGITAL IMAGE AUTHENTICATION
Svetozar Ilchev and Zlatoliliya Ilcheva

HOT TOPIC DETECTION BASED ON A TRIPARTITE MODEL
Pingbo Yuan, Bing Wang and Nenghai Yu

A NEW HYBRID METHOD FOR DATABASE SELECTION IN MULTI-DATABASE MINING
Aidin Davaran and Hassan Rashid

SCALING UP THE ACCURACY OF AVERAGED ONE-DEPENDENCE ESTIMATORS WITH DECISION TREE-BASED ATTRIBUTE WEIGHTED
Jia Wu, Zhihua Cai, Zhechao Gao and Yaodong Zhang

CHURN PREDICTION IN WIRELESS TELECOMMUNICATION - FUZZY DECISION TREES AND PATTERN TREES
Roland Merheb
ABSTRACT
Websites are typically designed attending to a variety of criteria. However, website structure determines browsing behavior and wayfinding results. The aim of this study is to identify the main profiles of websites structures by modeling web sites as graphs and considering several Social Network Analysis features. A case study based on eighty corporate Spanish Universities websites has been used for this purpose. Obtained results allow the categorization of website design styles and provide guidelines to assist designers to better identify areas for improvement and creation of effective Websites.

KEYWORDS
Website Structure, Link Analysis, Social Network Analysis, Factor analysis.

1. INTRODUCTION
The Web is an enormous set of documents connected through hypertext links created by designers of Web sites. Publishing on the Web is more than just setting up a page on a site; it also usually involves linking to other pages on the Web. The study of web links can offer a valuable source of information not only for developing informetric theory but also for studying link patterns between network entities (Yang & Qin, 2008). Link analysis methods can provide a quantitative measure about the quality of web pages (Bar-Ilan, J., 2005). In this sense, Social network analysis (SNA) has been frequently used for the study of link analysis (Park & Thelwall, 2003). SNA is a set of research procedures for identifying structures in social systems based on the relations among the system components, also referred to as nodes. In applying social network analysis methods to link analysis, websites or web pages are considered the actors, and therefore the nodes in the social network graph, while links are modelled as the relations between actors, representing the edges of the graph (Iacobucci, 1994). The purpose of this paper is to study websites structure patterns by modelling websites as connected graphs and by extracting several SNA features. Obtained results will highlight different websites profiles attending to their internal structure. This structure is closely related to users’ navigation experience. Badly designed Websites frustrate users and cause them to leave as they cannot find what they need. The reasons cited for the users’ negative experience include unavailability of information and, above all, difficulties for finding the required information. The rest of the paper is organized as follows. The next section analyzes the methodology. The case study based on eighty Spanish Universities is described in section 3, and the proposed methodology is applied in section 4 to obtain the websites structure patterns. Finally, the conclusions are withdrawn.

2. METHODOLOGY
Social network analysis arose from use of the mathematical model of graphs applied in the analysis of social relationships between actors (Wasserman and Faust 1994). Social network analysis may be viewed as a broading or generalization of standard data analytic techniques and applied statistics which usually focus on observational units and their characteristics (Wasserman & Faust, 1994; Toral et al., 2009a).
2.1 SNA Features of Websites

Networks representing web sites are collected starting at a given page (the root of the institutional web site) and then following the out links to other pages. Two different kinds of networks are considered for each web site. The first one is the domain network in which nodes represent sub domains or external domains different to the root domain. Arcs represent the link among them. The second network is the page network containing all the web pages of the institutional web site and the links among them. In the context of link analysis, the referred domain network is a star network. Several indicators related to its size have been measured in terms of nodes and lines. Finally, the density and average degree of the network have also been considered as indicators. Density is related to the number of lines and degree is a measure of the number of ties in which each vertex is involved. The referred page network is a more complex network, with a higher size and a much higher number of links than the domain network. Consequently, a higher number of social network features can be extracted.

- Size: the number of nodes is representing the number of web pages and arcs are representing the interrelations among these web pages. An important parameter to be chosen is the depth of link coverage when capturing web site information. A depth of seven has been used in this study.
- Density: density is defined as the number of lines in a simple network, expressed as a proportion of the maximum possible number of lines. A different measure of density is based on the idea of the degree of a node, which is the number of lines incident with it (Toral et al., 2009b). Finally, density can be measured alternatively suing an egocentric point of view; the egocentric density of a node is the density of ties among its neighbors (Nooy et al., 2005).
- Components: A strong component is a maximal strongly connected subnetwork. A network is said to be strongly connected if each pair of vertices is connected by a path, taking into account the direction of arcs (Nooy et al., 2005).
- K-cores: a k-core is a sub-network in which each node has k degree in that sub-network. The core with the highest degree is the central core of the network, detecting the set of nodes where the network rests on.
- Distance: it is defined as the number of steps in the shortest path that connect two nodes. In the case of web sites, there is a clearly defined main node which is the root of the network.
- Closeness centralization: it is an index of centrality based on the concept of distance. The closeness centrality of a node is calculated considering the total distance between one node and all other nodes, where larger distances yield lower closeness centrality scores (Toral et al., 2009c).
- Betweenness: it is a measure of centrality that rests on the idea that a person is more central if he or she is more important as an intermediary in the communication network (Nooy et al., 2005). It depends on the extent to which a node is needed as a link to facilitate the connection of nodes within the network.
- Partition correlation: A partition of a network is a classification or clustering of the nodes in the network such that each node is assigned to exactly one class or cluster. Two important partitions can be extracted: the k-neighbor partition, in which nodes are clustered using the distance to the root node, and the out-degree partition, in which nodes are clustered attending to their out-degree value. Two types of association indices are computed: Cramer’s V and Rajski’s information index (Nooy et al., 2005). Cramer’s V measures the statistical dependence between two classifications. Rajski’s indices measure the degree to which the information in one classification is preserved in the other classification.

3. CASE STUDY

The case study includes up to 80 Spanish University web sites. All of them are included in the Webometrics Ranking of World Universities (www.webometrics.org). They cover almost the whole range of webometrics Ranking, and exhibit a wide variety of sizes in terms of domains and web pages. More than 718000 web pages and more than four million outlinks have been considered through the analysis. Figure 1 and Figure 2 shows the particular case of the domain and page network of the University of Seville.
The social network features previously described have been measured, considering in some cases the whole network and in some cases a subnetwork. As a result, 26 indicators have been obtained (Table 1).

Table 1. List of selected indicators.

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Network</th>
<th>Description</th>
<th>Network</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>Density</td>
<td>Domain Network</td>
<td>Standard deviation of vertices betweeness centrality</td>
<td>Page network of (k)-cores, (k&gt;0)</td>
</tr>
<tr>
<td>I2</td>
<td>Number of pages</td>
<td>Page Network</td>
<td>Average value of vertices betweeness centrality</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I3</td>
<td>Total number of lines</td>
<td>Page Network</td>
<td>Standard deviation of vertices betweeness centrality</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I4</td>
<td>Density</td>
<td>Page Network</td>
<td>% of pages included in strong components</td>
<td>Page Network</td>
</tr>
<tr>
<td>I5</td>
<td>Average out-degree</td>
<td>Page Network</td>
<td>Average value of closeness centrality</td>
<td>Page Network</td>
</tr>
<tr>
<td>I6</td>
<td>Number of pages in the last level</td>
<td>Page Network</td>
<td>Standard deviation of closeness centrality</td>
<td>Page Network</td>
</tr>
<tr>
<td>I7</td>
<td>Standard deviation of out-degree</td>
<td>Page Network</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I8</td>
<td>Number of pages (excluding vertices with out-degree=0)</td>
<td>Page Network</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I9</td>
<td>Density (excluding vertices with out-degree=0)</td>
<td>Page Network (excluding vertices with out-degree=0)</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I10</td>
<td>Average degree</td>
<td>Page Network (excluding vertices with out-degree=0)</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I11</td>
<td>Average out-degree</td>
<td>Page Network (excluding vertices with out-degree=0)</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I12</td>
<td>Standard deviation of closeness centrality</td>
<td>Page Network (excluding vertices with out-degree=0)</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
<tr>
<td>I13</td>
<td>Number of vertices with betweeness centrality &gt; 0</td>
<td>Page network (excluding vertices with out-degree=0)</td>
<td>Cramer's V</td>
<td>Page network (excluding vertices with out-degree=0)</td>
</tr>
</tbody>
</table>

4. RESULTS

Factor analysis has been applied to categorize websites according to the style in which they have been designed. Factor analysis is a data reduction technique used to find homogeneous groups in a large set of
data. These groups represent the underlying variables or factors, which can explain the pattern of correlations within a set of observed variables (Rencher, 2002). In factor analysis it is usual to consider a number of factors able to account for more than 70% of the total sample variance. In our case study, this value is reached with four factors. Using the associated eigenvectors, factor loadings can be estimated. Sometimes, it is difficult to perform the right interpretation of factors using the estimated loadings. Fortunately, factor loading can be rotated through the multiplication by an orthogonal matrix (Varimax rotation). The rotated loadings preserve the essential properties of the original loadings. Typically, a loading threshold value of 0.6 is usually considered (Rencher, 2002). The resulting aggregation of variables leads to the identified latent factors of Table 2.

Table 2. Identified factors

<table>
<thead>
<tr>
<th>Factor 1: Highly structured websites</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 Density (domain network)</td>
<td>.731</td>
</tr>
<tr>
<td>14 Density (page network)</td>
<td>.617</td>
</tr>
<tr>
<td>19 Density (page network excluding out-degree=0)</td>
<td>.728</td>
</tr>
<tr>
<td>12 Standard deviation of closeness centrality</td>
<td>.659</td>
</tr>
<tr>
<td>14 Standard deviation of vertices betweeness centrality</td>
<td>.840</td>
</tr>
<tr>
<td>15 Average value of vertices betweeness centrality</td>
<td>.808</td>
</tr>
<tr>
<td>16 Standard deviation of vertices betweeness centrality</td>
<td>.800</td>
</tr>
<tr>
<td>17 Cramer's V</td>
<td>.716</td>
</tr>
<tr>
<td>18 Rajski (C1&lt;-&gt;C2)</td>
<td>.750</td>
</tr>
<tr>
<td>122 Cramer's V</td>
<td>.683</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 2: centralized websites</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>19 % of pages included in strong components</td>
<td>.748</td>
</tr>
<tr>
<td>120 Average value of closeness centrality</td>
<td>.930</td>
</tr>
<tr>
<td>121 Standard deviation of closeness centrality</td>
<td>.785</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 3: Large websites</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 Number of pages</td>
<td>.885</td>
</tr>
<tr>
<td>13 Total number of lines</td>
<td>.912</td>
</tr>
<tr>
<td>15 Number of pages in the last level</td>
<td>.779</td>
</tr>
<tr>
<td>18 Number of pages</td>
<td>.841</td>
</tr>
<tr>
<td>113 Number of vertices with betweeness centrality &gt; 0</td>
<td>.873</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Factor 4: Partitioned websites</th>
<th>Factor loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 Average out-degree</td>
<td>.726</td>
</tr>
<tr>
<td>17 Standard deviation out-degree</td>
<td>.624</td>
</tr>
<tr>
<td>110 Average degree</td>
<td>.878</td>
</tr>
<tr>
<td>111 Average out-degree</td>
<td>.715</td>
</tr>
<tr>
<td>124 Egocentric density (average value)</td>
<td>.687</td>
</tr>
<tr>
<td>123 Egocentric density (average value)</td>
<td>.713</td>
</tr>
</tbody>
</table>

Extraction Method: Principal Component Analysis
Rotation Method: VARIMAX with Kaiser Normalization
The rotation has converged in 6 iterations

On the other hand, factor scores are used to categorize the original sample of Universities, which can be approximated to one of the identified latent factors. An analysis of variance (ANOVA) has been performed to check the null hypothesis of equal population means, which have been rejected in all the cases with a significance value below 0.05. Using the information of the factor loadings as well as the mean values of the categorized groups of Universities, the following websites structure patterns can be highlighted:

- Factor 1 represents highly structured websites. The high value of Rajski and Cramer's V information indices indicates the out-degree is growing as vertices are more distant from the root domain. The high value of average value and standard deviation of vertices betweeness centrality suggest the website is structured through highly interconnected vertices spread over the website, following a certain tree structure. Finally, factor 1 exhibits a high value of density due to the fact of being small web sites as compared to the websites assigned to other factors.

- Factor 2 represents a more centralized structure in the sense of distance to the root domain. There is a core of highly interconnected pages around the root domain, facilitating the accessibility of information. Website is organized in a flat structure as compared to rest of factors.

- Factor 3 represents large websites, which probably have been growing during the years in a certain chaotic progression. The number of pages grows geometrically with the depth level, so it is necessary a long navigation process to achieve the desired information. Most of web pages play a betweeness role as
there is not a formal structure under which the web site was designed.

- Finally, factor 4 represents partitioned web sites where the global network could be considered as the sum of more or less independent subnetworks. In this case, websites are organized around subdomains related to different areas of the organization.

Basically, identified profiles of web site structures respond to two basic strategies when deciding their final structure (Tan and Wei, 2006). The first strategy consists of offering a structure which makes sense to the final user. In this sense, web sites sacrifices accessibility of information looking for a more structured navigation scheme. Factors 1, 3 and 4 could be included in this strategy. The alternative option consists of reducing big structures under the assumption that user performance is optimal when breadth and depth of Website is kept to a moderate level (Tan and Wei, 2006). This is the strategy represented by factor 2.

5. CONCLUSION

This paper proposes the identification of web structure patterns using SNA techniques. As a case study, SNA features from eighty institutional websites corresponding to Spanish Universities have been extracted and statistically analyzed. Results distinguish four types of websites organization according to their structure. Three of them exhibit different kinds of structured organization while the last is closer to a flat organization, to emphasize the accessibility of information. Although the study is restricted to Spanish Universities, it could be extended to Universities all over the world, or even to different corporate websites as a future work.

ACKNOWLEDGEMENT

This work has been supported by the Spanish Ministry of Education and Science (Research Project with reference DPI2007-60128) and the Consejería de Innovación, Ciencia y Empresa (Research Project with reference P07-TIC-02621).

REFERENCES


