STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVIII**, Number 4, 2013

ANALYSING WEB USAGE WITH FORCE-DIRECTED GRAPHS

SANDA MARIA DRAGOŞ AND ALINA MIHAELA BELDEAN

Abstract. This paper presents an analysis of web usage data by using force-directed graphs as a visualization instrument. Web analytics instruments provide useful insights but they lack the visual representation of referrer-referred links between different web pages.

The paper presents a work-in-progress investigation that draws its approach from the way data are nowadays analysed by visual representation. Using this approach for this particular type of data is (to the best of the author's knowledge) innovative.

Graphs are widely used to visualize data structured as objects and the relationships between them. Typically they are used to represent relationships between individuals [14] (as in social networks), traffic between distinct location [6] (in traffic networks) or relationships between genes [23] (in genetics).

This paper presents the attempt to analyse web usage data using graph visualization.

1. Force-Directed Graphs as a Visualization Instrument

Graphs are generally drawn as node-link diagrams in which the vertices are represented as disks or boxes and the edges are represented as line segments. The most well-known class of node-link-based visualization techniques for general graphs is the class of force-directed methods and its derivatives [13, 19]. The usefulness of the graph representation is dependent on the aesthetics of the drawing. Therefore, there are many studies that tried to improve the graph aesthetics by minimizing the edge crossing, evenly distributing vertices, and the depiction of graph symmetry [18].

Force-directed methods consider graph nodes so that all edges are of more or less the same length. To reduce the edge crossing, there are assigned forces

Received by the editors: October 29, 2013.

2010 *Mathematics Subject Classification.* 68T30 Knowledge representation, 68P20 Information storage and retrieval.

1998 *CR Categories and Descriptors.* H.3.5 [**Information Systems**]]: Information Storage and Retrieval – *On-line Information Services.*

Key words and phrases. web usage, force-directed graph, visualization.

that make edges to behave like springs and the nodes to behave like electrically charged particles. Therefore, the edges have forces that attract the nodes, while the nodes have forces that reject all other nodes. The entire graph is then simulated like a physical system. Computational complexity and layout stability issues emerged. They were treated by various approaches [20, 16, 21].

There are a multitude of software, systems and providers of systems for drawing graphs. Some of them are:

- **Cytoscape**: an open-source software for visualizing molecular interaction networks [26];
- **Gephi**: an open-source network analysis and visualization software [3];
- **Graphviz**; an open-source graph drawing system from AT&T Corporation [12];
- **Mathematica**: a general purpose computation tool that includes 2D and 3D graph visualization and graph analysis tools [1];
- **Microsoft Automatic Graph Layout**: a .NET library (formerly called GLEE) for laying out graphs [25];
- **Tom Sawyer Software**: Tom Sawyer Perspectives is a graphics-based software for building enterprise-class data visualization and social network analysis applications. It is a Software Development Kit (SDK) with a graphics-based design and preview environment [22];
- **Tulip**: (software) [2];
- **yEd**: a widely used graph editor with graph layout functionality [27].

There are also a number of JavaScript libraries deploying force-directed graph layout algorithms. Some of them are:

- **Springy.js**: a force directed graph layout algorithm in JavaScript [17];
- **Protovis**: a free and open-source (provided under the BSD License) graphical toolkit for visualization. It uses JavaScript and SVG for web-native visualizations. [5]. The Protovis team is now developing a new visualization library, D3.js, with improved support for animation and interaction.
- **D3.js**: a JavaScript library for visualization of data [4];
- **Graph JavaScript framework**: is a freely distributable library under the terms of an MIT-style license [15]. The algorithm is based on a spring-style layouter of a Java-based social network tracker PieSpy written by Paul Mutton [24].

2. Graph JavaScript framework

This paper presents such a study by using the Graph JavaScript framework, version 0.0.1 [24]. This instrument positions each node from the graph in the origin of the coordinate system and then it computes the *repulsiveForce* and the *attractiveForce* as depicted in Listings 1 and 2 to rearrange the nodes for a optimal visualization.

At first all nodes are placed on $(0, 0)$ coordinates, after which on each node are applied node-node repulsion forces and edge attraction forces to reposition them. In order to compute the repulsion force between two nodes, the square of the distance between those nodes (i.e. $d2 = d^2$) is computed. If this distance is less than 0.1 (i.e. $d^2 < 0.01$ as described in Listings 1, line 5) a new, random distance will be considered for recomputing the new positions of those nodes. The repulsive force will modify the position of the current nodes only if it is below a predefined threshold called *maxRepulsiveForceDistance*.

Listing 1. The implementation of the *repulsiveForce*

```
 1  layoutRepulsive: function(node1, node2) {
        var dx = node2.layoutPosX - node1.layoutPosX;
 3      var dy = node2.layoutPosY - node1.layoutPosY;
        var d2 = dx * dx + dy * dy;
 5      if(d2 < 0.01) {
            dx = 0.1 * Math.random() + 0.1;
 7          dy = 0.1 * Math.random() + 0.1;
            var d2 = dx * dx + dy * dy;
 9      }
        var d = Math.sqrt(d2);
11      if(d < this.maxRepulsiveForceDistance) {
            var repulsiveForce = this.k * this.k / d;
13          node2.layoutForceX += repulsiveForce * dx / d;
            node2.layoutForceY += repulsiveForce * dy / d;
15          node1.layoutForceX -= repulsiveForce * dx / d;
            node1.layoutForceY -= repulsiveForce * dy / d;
17      }
    }
```

In this case (i.e., if the condition on line 11 from Listings 1 is true) the *repulsiveForce* is established by the equation (1), where k is a constant.

$$repulsiveForce = \frac{k^2}{d} \tag{1}$$

In the case of attractive forces exercised by an edge, the same procedure is done by determining the value of the distance d between the nodes connected by the current edge, as depicted in Listings 2. However, in the case in which d exceeds the established threshold (i.e., *maxRepulsiveForceDistance*), d will be considered to be this threshold.

Listing 2. The implementation of the *attractiveForce*

```
   layoutAttractive: function(edge) {
 2     var node1 = edge.source;
       var node2 = edge.target;

       // the d2 computation phase (as previously)
 6
       var d = Math.sqrt(d2);
       if(d > this.maxRepulsiveForceDistance) {
           d = this.maxRepulsiveForceDistance;
10         d2 = d * d;
       }
       var attractiveForce = (d2 - this.k * this.k)/this.k;
       if(edge.weight == undefined || edge.weight < 1)
14         edge.weight = 1;
       attractiveForce *= Math.log(edge.weight) * 0.5 + 1;

       node2.layoutForceX -= attractiveForce * dx / d;
18     node2.layoutForceY -= attractiveForce * dy / d;
       node1.layoutForceX += attractiveForce * dx / d;
       node1.layoutForceY += attractiveForce * dy / d;
   }
```

The computation of the *attractiveForce* is established by the formula depicted in equation (2), where *edge.weight* ≥ 1 (see lines 13-14 from Listings 2).

$$attractiveForce = \left(\frac{d^2}{k} - k\right) \times \left(\frac{\log edge.weight}{2} + 1\right) \tag{2}$$

If the *edge.weight* is undefined or less that unity, it is considered to be unity. Therefore, supra-unitary edge-weighted graphs are also processed by the *Graph JavaScript framework*.

3. Force-Directed Graphs for Web Usage Analytics

Graph visualization is used to interpret a large range of data. However, web usage data is not yet (to our best knowledge) interpreted using such instruments. Typically, web usage data are interpreted using web analytics instruments. However, there are relationship-based data that would benefit from a graph visualization. One example is the referrer-referred relationship between the visited web pages. The sites used for this analysis are:

- **http://www.cs.ubbcluj.ro/∼sanda** the author's website containing *personal*, but mostly *teaching* and *research* information.

- **PULSE** a PHP Utility used in Laboratories for Student Evaluation [7, 8, 9]. The web usage of PULSE was also analysed using a Web Analytics [10] and Formal Concept Analysis [11].

The graph is constructed using referrer-referred pairs of web-pages to describe an edge between two nodes as depicted in Figure 1.

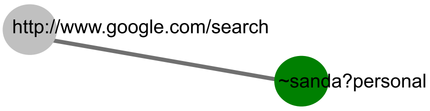


FIGURE 1. Edges as links between referrer-referred pairs of nodes

The web usage data used for these tests are the same as in the Formal Concept Analysis presented in paper [11]. This was done intentionally in order to have two perspectives in interpreting the same data set. Therefore, the analysis was performed on the data collected from the two months of the last academic year (i.e. April and May of 2012). The data, however, was pruned by the entries that did not contained a referrer. The first test result for the entire data set (i.e., **PULSE** and ∼**sanda** web accesses) is presented in Figure 2. There were a significant number of accesses on **PULSE** presented as the circular conglomeration of *blue* nodes from the bottom-left corner of the graph represented in this Figure. That is because **PULSE** is used on regular bases for the work within laboratories and for consulting the teaching related material on *anytime*, *anywhere* bases. The *green nodes* represent ∼**sanda** pages as either the web accessed pages or the referrer pages for other **PULSE** or ∼**sanda** pages. The *yellow* nodes are referrers from the web server of our department (i.e., starting with http://www.cs.ubbcluj.ro but different from **PULSE** and ∼**sanda** pages). The *red* nodes are all others referrers (i.e., from other sources).

For the next test there were considered only ∼**sanda** pages visited. The results are presented in Figure 3. As this graph contained less nodes and therefore it is more readable an arrow point was introduced to denote the role of the *referrer* or *referred* of a node.

Here it can be observed that the ∼**sanda?interests** pages form distinct formations which are referred directly by external nodes.

There are also *google* queries which leaded to ∼**sanda** pages, denoted by *gray* nodes, which refer the ∼**sanda?personal** and ∼**sanda?photos** pages. The *blue* (i.e., PULSE) nodes are all referrers to ∼**sanda** node-page. This node is also the most referred node being placed in the middle of its formation. That may be also because this page in the *homepage* of the site. Other important

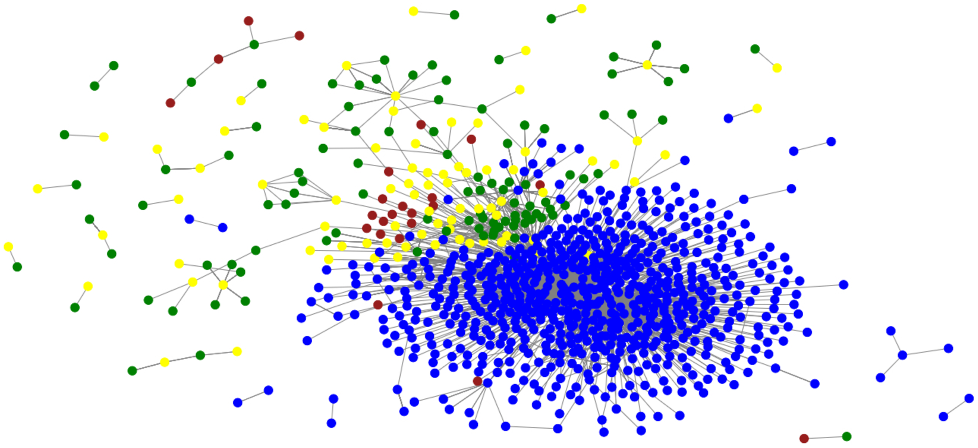
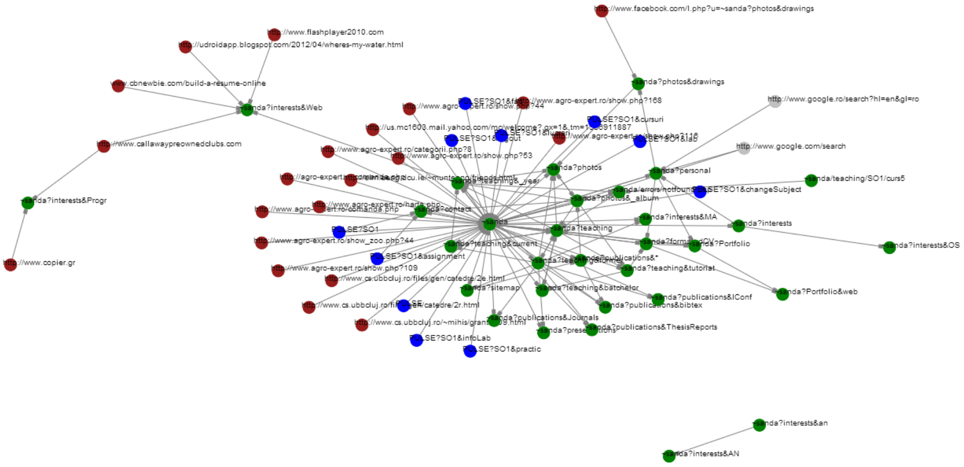


Figure 2. All logged data that had a referrer



http://www.facebook.com/l.php?u=~sanda?photos&drawings
http://www.flashplayer2010.com
http://udroidapp.blogspot.com/2012/04/wheres-my-water.html
www.cbnewbie.com/build-a-resume-online
~sanda?interests&Web
http://www.callawaypreownedclubs.com
~sanda?interests&Progr
http://www.copier.gr
~sanda?photos&drawings
http://www.agro-expert.ro/show.php?168
http://www.google.ro/search?hl=en&gl=ro
PULSE?SO1&cursuri
http://www.google.com/search
~sanda?photos
~sanda?personal
~sanda/teaching/SO1/curs5
PULSE?SO1&changeSubject
~sanda?photos&_album
~sanda?interests&MA
~sanda?interests
~sanda?interests&OS
~sanda?teaching
~sanda?Portfolio
~sanda?teaching&current
~sanda?teaching&tutorlat
~sanda?Portfolio&web
~sanda?sitemap
~sanda?teaching&batchelor
~sanda?publications&IConf
~sanda?publications&bibtex
~sanda?publications&Journals
~sanda?publications&ThesisReports
PULSE?SO1&infoLab
PULSE?SO1&practic
PULSE?SO1&assignment
PULSE?SO1
~sanda
~sanda?contact
http://www.agro-expert.ro/show_zoo.php?44
http://www.agro-expert.ro/show.php?109
http://www.cs.ubbcluj.ro/files/gen/catedre/2e.html
http://www.cs.ubbcluj.ro/~mihis/grant09.html
http://www.agro-expert.ro/categorii.php?8
http://www.agro-expert.ro/show.php?53
http://www.agro-expert.ro/harta.php
~sanda?interests&an
~sanda?interests&AN

Figure 3. ∼**sanda** referred data

nodes which are often referred are ∼**sanda?teaching**, ∼**sanda?publications**, ∼**sanda?contact** and ∼**sanda?photos**.

For the next set of tests there were considered the **PULSE** referred pages. As there were too many such nodes, the **PULSE** related data was split based on the following contexts:

- PULSE accesses before or after the login session;
- PULSE accesses during the login session:

– PULSE used by students;
– PULSE used by teacher.

Based on the tests done on the same set of data by using Formal Concept Analysis, students used PULSE in proportion of almost 40%, which in this case means 8334 accesses. The teacher using PULSE accessed it 434 times (=2.08%). The rest of 58.07% accesses (i.e., 12142 accesses) appear under the *'no login'* label.

The results from the PULSE accesses done before and after the login session (i.e., for which there is no *login* value) are depicted in Figure 4.

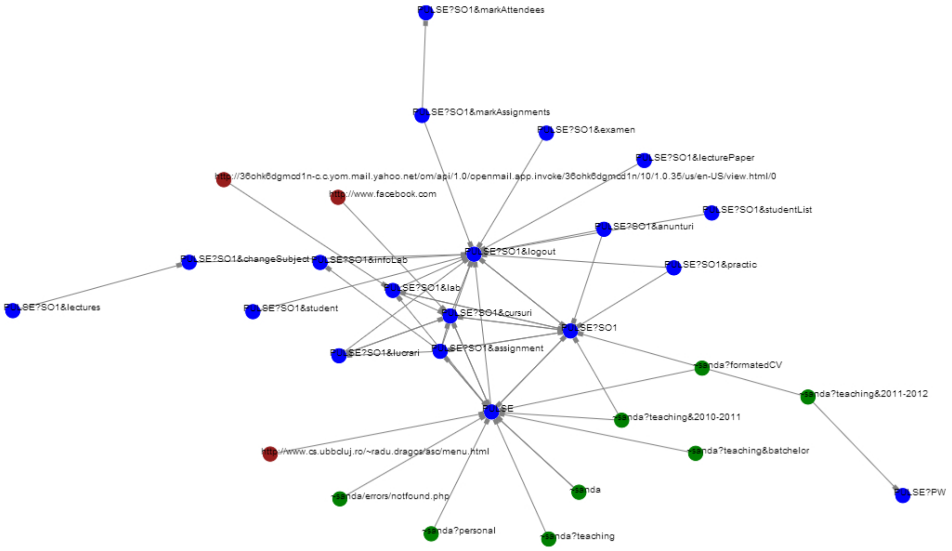


Figure 4. PULSE referred data before or after the login phase

Here it can be observed that the most accesses without a login are done on the **PULSE** homepage or after the PULSE actors have accessed the PULSE's logout functionality (i.e., a **PULSE?*&logout** page). Other PULSE pages were also accessed either when the session expired or if it was accessed a direct link to a PULSE page from an external page. The latter is the case of the **PULSE?SO1** or **PULSE?PW** page which had a direct link from the ∼*sanda?teaching* page. In all these situations the users are directed to the login **PULSE** page.

The next tests were performed on the PULSE data collected after PULSE's actors successfully passed the login phase. This results are shown in Figure 5. The most referred page was the **PULSE** homepage, which was also the most referred by sites outside PULSE. The majority of those pages were ∼*sanda?teaching* ones. However, direct links to the Operating Systems lecture material (i.e., **PULSE?SO1&cursuri**) were made also from *facebook* and *mail.yahoo*.

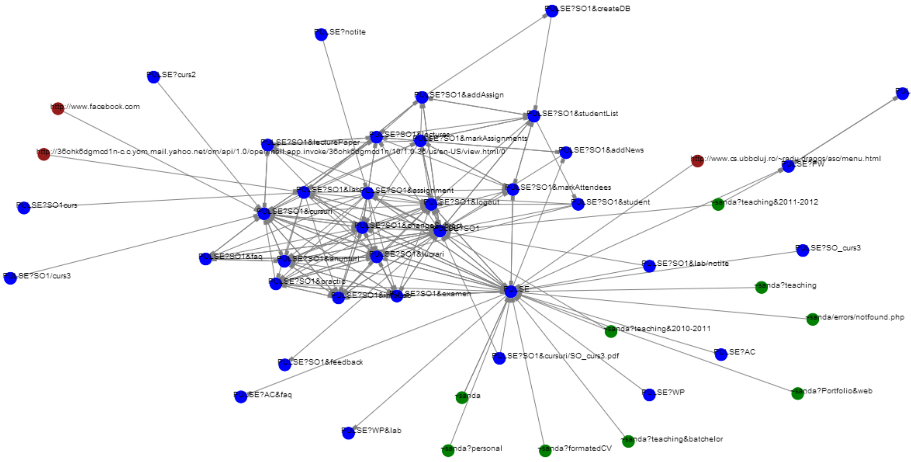


Figure 5. PULSE referred data during the login session

Other PULSE pages that were often referred relate to the Operating System (i.e. *SO1*) taught subject. As in the recorded semester (e.g. from February to June 2012), the subject studied was SO1 (i.e., Operating Systems) this result is justifiable. The Formal Concept Analysis gave the number of 7015 accesses. What is interesting to observe here is that the other subjects (i.e., AC, WP, PW), although studied in the previous semester are still revisited. This is surprising because students were examined on those subjects and they still return to revisit the information posted there.

The next phase in our testing is to split these PULSE accesses based on its actors. Therefore, Figure 6 shows the PULSE accesses done by students.

Here, as in the Formal Concept Analysis investigation, we wanted to check how well PULSE performs for the task that it was designed for: to see if students access the information provided. The distributions presented here confirmed our expectations. PULSE was designed as support instrument for laboratories and lectures. Therefore, the most accessed page was the **PULSE** homepage, with a percentage of accesses obtained from the Formal Concept Analysis of 47.64%. This page loads just after the login phase and contains general information for students, such as:

- the name of the authenticated person;

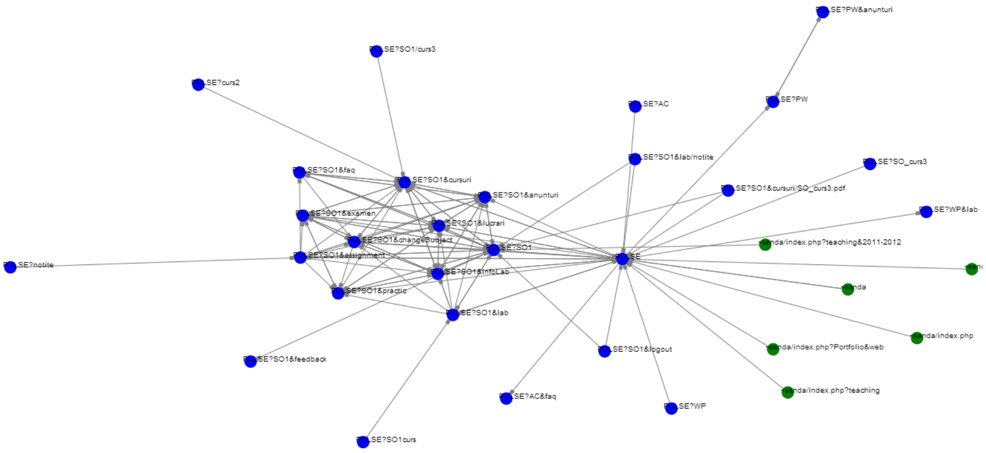


Figure 6. PULSE used by students

- the group of the student;
- notifications/announcements from the teacher;
- for each laboratory specific information such as:
  - the week within the semester and corresponding calendar dates;
  - the name of the concept studied;
  - the corresponding assignment reference;
  - the mark if the assignment was handed;
  - and the attendance status
- lab activity (i.e., average score and the total number of attendances)
- the marks for the practical and written exam (at the end of the semester) as well as the final mark.

Therefore, this page is visited very often. There are other facilities offered by PULSE which can be depicted in Figure 6. The **PULSE?*&anunturi** page, which was accessed by students (as determine by the FCA) 69 times. This page which contains all notifications/announcements made by the teacher for that specific subject. The last announcement is always posted also on the *'PULSE homepage'*. The *'faq'* (Frequently Asked Questions) page was accessed 22 times, while only one student access was made to send a *feedback*. All these pages are at the edge of the graph, being scarcely accessed, compared with the highly connected pages from the centre of the graph.

As also shown in Figure 6, the most visited lecture related pages are those containing the theoretical support (i.e., **PULSE?SO1&cursuri**). Then, there are the test papers (i.e., presented as the **PULSE?SO1&lucrari** node) during lectures and their results (including statistics and explanation how their marks will help the student). Similar with lectures, the **PULSE?*&lab** pages contain the technical support presenting the required concepts and examples. To better understand the concepts, students are given *'assignments'* represented by the **PULSE?SO1&assignmet** node.

PULSE facilities offered for teachers are presented in Figure 7.

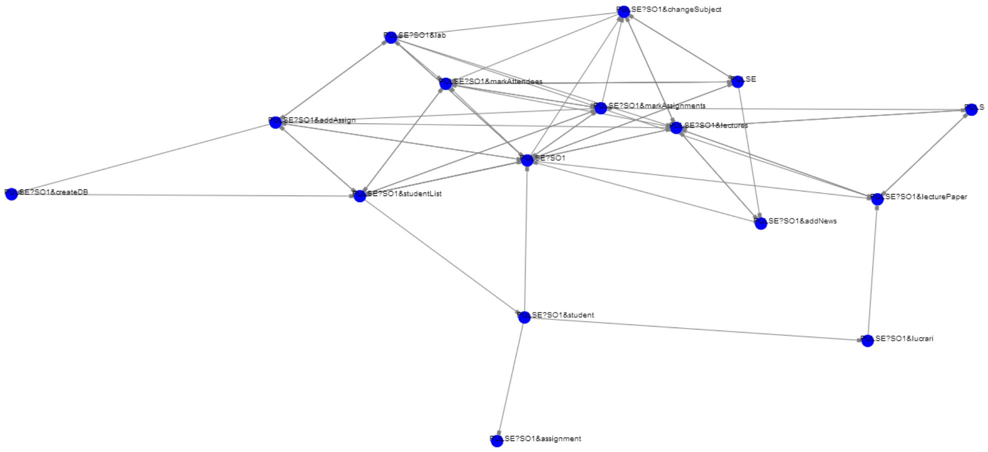


Figure 7. PULSE used by the teacher

The main PULSE facilities, in their most accessed order are:

- **markAssignments** - is to mark student scores (the most used PULSE facility, in the proportion of 39.40% according to the FCA analysis);
- **lectures** - check the course support (in the proportion of 11.29% according to the FCA analysis);
- **markAttendees** - mark student attendances (in the proportion of 11.06% according to the FCA analysis);
- **labs** - check lab support (in the proportion of 5.07% according to the FCA analysis);
- **addAssign** - assign random tasks for students (in the proportion of 4.38% according to the FCA analysis);
- **studentList** - list all students with their marks, attendances and final scores (in the proportion of 3.00% according to the FCA analysis);
- **lecturePaper** and **lucrari** - list all students which have lecture paper marks (in the proportion of 1.61% according to the FCA analysis);
- **addNews** - add notifications/announcements (in the proportion of 0.92% according to the FCA analysis).

There are also other PULSE pages, not regarded by FCA: **changeSubject** page, which is also considerably accessed.

4. Conclusions and Future Work

The insights brought by the new way of investigation of web usage data allowed to visualize the way in which pages are visited, from where and in which order. Also, the results showed that the most visited and highly linked pages were placed in the centre of the forced-directed graph while the other pages were placed on the edge of the graph. These results also concurred with the ones obtained by analysing the same set of data with the Formal Concept Analysis. These results were published in [11].

For a further research, we would like to investigate the forced-directed graphs extended by the use of some of the social networks measures such as *betweenness*, *centrality* and *clustering*.

References

[1] *GraphPlot - Wolfram Mathematica 9 Documentation.* http://reference.wolfram.com/mathematica/ref/GraphPlot.html. [Online; accessed 10-April-2013].

[2] D. Auber, *Tulipa huge graph visualization framework*, in Graph Drawing Software, Springer, 2004, pp. 105–126.

[3] M. Bastian, S. Heymann, and M. Jacomy, *Gephi: An open source software for exploring and manipulating networks*, in International AAAI conference on weblogs and social media, vol. 2, AAAI Press Menlo Park, CA, 2009.

[4] M. Bostock, *Data-driven documents (d3.js), a visualization framework for internet browsers running javascrip.* http://d3js.org/, 2012. [Online; accessed 10-April-2013].

[5] M. Bostock and J. Heer, *Protovis: A graphical toolkit for visualization*, Visualization and Computer Graphics, IEEE Transactions on, 15 (2009), pp. 1121–1128.

[6] H. Djidjev, G. Sandine, C. Storlie, and S. Vander Wiel, *Graph based statistical analysis of network traffic*, in Proceedings of the Ninth Workshop on Mining and Learning with Graphs, 2011.

[7] S. Dragos, *PULSE - a PHP Utility used in Laboratories for Student Evaluation*, in International Conference on Informatics Education Europe II (IEEII), Thessaloniki, Greece, November 2007, pp. 306–314.

[8] ———, *PULSE Extended*, in The Fourth International Conference on Internet and Web Applications and Services, Venice/Mestre, Italy, May 2009, IEEE Computer Society, pp. 510–515.

[9] ———, *Current Extensions on PULSE*, Studia Universitatis Babes-Bolyai Series Informatica, LV (2010), pp. 51–60.

[10] S. Dragos and R. Dragos, *WATEC: a Web Analytics Tool for Educational Content*, KEPT2009 Knowledge Engineering Principles and Techniques, Selected Papers (2009), pp. 320–327.

[11] S. Dragos and C. Sacarea, *Analysing the Usage of Pulse Portal with Formal Concept Analysis*, Studia Universitatis Babes-Bolyai Series Informatica, LVII (2012), pp. 65–75.

[12] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull, *Graphviz and dynagraphstatic and dynamic graph drawing tools*, in Graph Drawing Software, Springer, 2004, pp. 127–148.

[13] T. M. Fruchterman and E. M. Reingold, *Graph drawing by force-directed placement*, Software: Practice and experience, 21 (1991), pp. 1129–1164.

[14] D. Greene and P. Cunningham, *Producing a unified graph representation from multiple social network views*, arXiv preprint arXiv:1301.5809, (2013).

[15] A. Hellesoy and D. Hoover, *Graph javascript framework, version 0.0.1.* http://snipplr.com/view/1950/graph-javascript-framework-version-001/, 2006. [Online; accessed 10-April-2013].

[16] I. Herman, G. Melançon, and M. S. Marshall, *Graph visualization and navigation in information visualization: A survey*, Visualization and Computer Graphics, IEEE Transactions on, 6 (2000), pp. 24–43.

[17] D. Hotson, *Springy - A force directed graph layout algorithm in JavaScript.* http://getspringy.com/, 2010. [Online; accessed 10-April-2013].

[18] Y. Hu, *Efficient, high-quality force-directed graph drawing*, Mathematica Journal, 10 (2005), pp. 37–71.

[19] T. Kamada and S. Kawai, *An algorithm for drawing general undirected graphs*, Information processing letters, 31 (1989), pp. 7–15.

[20] D. H. Y. Koren, *A fast multi-scale method for drawing large graphs*, Journal of graph algorithms and applications, 6 (2002).

[21] Y. Koren, L. Carmel, and D. Harel, *Drawing huge graphs by algebraic multigrid optimization*, Multiscale Modeling & Simulation, 1 (2003), pp. 645–673.

[22] B. Madden, P. Madden, S. Powers, and M. Himsolt, *Portable graph layout and editing*, in Graph Drawing, Springer, 1996, pp. 385–395.

[23] D. Merico, D. Gfeller, G. D. Bader, et al., *How to visually interpret biological data using networks*, Nature biotechnology, 27 (2009), p. 921.

[24] P. Mutton, *Inferring and visualizing social networks on internet relay chat*, in Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on, IEEE, 2004, pp. 35–43.

[25] L. Nachmanson, G. Robertson, and B. Lee, *Drawing graphs with glee*, in Graph Drawing, Springer, 2008, pp. 389–394.

[26] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, *Cytoscape 2.8: new features for data integration and network visualization*, Bioinformatics, 27 (2011), pp. 431–432.

[27] R. Wiese, M. Eiglsperger, and M. Kaufmann, *yfilesvisualization and automatic layout of graphs*, in Graph Drawing Software, Springer, 2004, pp. 173–191.

Department of Computer Science, Faculty of Mathematics and Computer Science, Babeş-Bolyai University, 1 M. Kogălniceanu St., 400084 Cluj-Napoca, Romania

E-mail address: `sanda@cs.ubbcluj.ro`