STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LVII**, Number 3, 2012

ANALYSING THE USAGE OF PULSE PORTAL WITH FORMAL CONCEPT ANALYSIS

SANDA DRAGOŞ AND CHRISTIAN SĂCĂREA

Abstract. This paper aims to present an analysis of web usage using Conceptual Knowledge Processing and Representation tools, mainly Formal Concept Analysis and the knowledge management system TOSCANAJ. We describe several knowledge maps in form of conceptual hierarchies in order to highlight how these tools might be used for analysing access data in the PULSE system.

1. Introduction

Formal Concept Analysis (FCA) [3] is a powerful tool in order to represent and process knowledge. We have been mainly interested in evaluating and representing web usage data, which are usually interpreted using web analytics or data mining techniques.

By this approach, data have been gathered and then analysed by the knowledge management system TOSCANA. Using this system, concepts reflecting web usage have been built and then represented in conceptual hierarchies. These hierarchies are then used as knowledge maps allowing us to analyse and hence to build valuable judgements over our data.

To the best of our knowledge FCA was not used for interpreting web site usage data.

2. Formal Concept Analysis

Formal Concept Analysis is a mathematical theory widely used in data analysis. The basic structure is the *formal context*, which exploits the fact that data is quite often represented as objects and attributes. Attributes

Received by the editors: October 6, 2012.

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. formal concept analysis, web analytics.

might be either simple (attributes have binary values, YES/NO) or many-valued (attributes have values). Objects are linked to their attributes by an incidence relation. Because of the specificity of our data, we have considered manyvalued contexts. These are then scaled, i.e., transformed by means of a conceptual analysis into simple, binary contexts. For these contexts, concept are built and displayed in conceptual hierarchies, i.e., order diagrams.

Formally, a formal context is a triple (G, M, I), where G is a set of objects, M is a set of attributes, and I binary relation $I \subseteq G \times M$, called the *incidence relation*. Thus, gIm is read *the object g has attribute m*. A fininte context can be represented as a cross-table, the rows labeled by object names, the columns by attribute name and the incidence relation is represented by crosses.

A manyvalued context is defined as (G, M, W, I), G being the set of object, M the set of attributes, W the set of attribute values, and I a ternary relation linking object, attributes and their values, also called incidence relation. Here $(g, m, w) \in I$ means that *object g has attribute m with value w*. However, the data needs to be interpreted/converted from many-valued attributes into a single value attribute. This process is called conceptual scaling by Ganter and Wille [3].

FCA is a formalization of the classical, philosophical understanding of concepts and their importance for knowledge. We define two derivation operators, which proves to be a Galois connection between the power sets of G and M (see [3]). These operators are defined as follows: Let (G, M, I) be a formal context and $A \subseteq G, B \subseteq M$ be subsets of objects and attributes, respectively.

$$A' := \{m \in M \mid \ gIm, \ \forall g \in A\}.$$
$$B' := \{g \in G \mid \ gIm, \ \forall m \in B\}.$$

A formal concept is a pair (A, B) of sets, $A \subseteq G$, $B \subseteq M$ with $A' := B, B' := A$. A descriptive interpretation is the following: All commonly shared attributes of all objects in A belong to B and all attributes in B are shared by all objects in A. Not so obvious is the fact that (A, B) is maximal with this property. Using the cross-table representation of data, a concept is a maximal rectangle of crosses, property which is directly derived from the above definition. Concepts might also be understood as basic units of knowledge, since they reflect a basic structure of data clustering.

We denote the set of all concepts by $\mathfrak{B}(G, M, I)$. This set is structured by an order relation, called the *subconcept-superconcept relationship*. This is a specialization-generalization relationship for concepts. We say that (A, B) is a *subconcept* of (C, D) (or a specialization of it) if and only if A is a subset of C (which is of course equivalent to D being a subset of B). The concept (C, D) is called superconcept (or generalization) of (A, B). This relation is an order relation over the set of all concepts:

$$(A, B) \leq (C, D) :\Leftrightarrow A \subseteq C (\Leftrightarrow D \subseteq B).$$

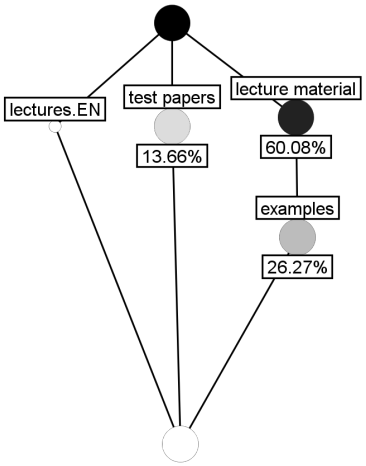

FIGURE 1. Example of a Hasse diagram representing conceptual hierarchies

The ordered set $(\mathfrak{B}(G, M, I), \leq)$ is a complete lattice, called *conceptual hierarchy*, or *concept lattice* (for more information we refer to [3]). The conceptual hierarchy is then used as a knowledge map, since it represents the encoded knowledge in the data under analysis.

These conceptual hierarchies can be graphically represented by a order diagrams with a simple and easy to read structure and labelling as depicted in Figure 1.

How do we read such a diagram? Every node is a (formal) concept, i.e., a maximal collection of objects and common attributes. The set of objects is called extent, that of attributes intent. The same object or attribute might be part of different concepts, thus establishing a hierarchy which enables navigation. To illustrate this of Figure 6, the node labelled *'examples'* lies below the node labelled *'lecture material'*, that means that *'examples'* is a subconcept of *'lecture material'*. The percentages below every node represent the access percentages.

R. Wille considered ([4]) that knowledge can be represented in conceptual landscapes, hence navigation must be provided from one concept to another. The conceptual hierarchies, as being described in [5] and [6], act both as knowledge maps but also as a guide through the data set under consideration. They not only represent knowledge, but also make possible knowledge acquisition, knowledge processing and a special view, called *conceptual*. Conceptual means that the queries we make, the navigation and zooming into different parts of our knowledge landscapes are driven by a set of concepts, a set of

clearly structured ideas which can also be read off from the hierarchies under consideration.

This view has been implemented in the knowledge management system ToscanaJ [1]. This system comprises three different tools, Elba, Siena and Toscana itself.

Data is usually stored in a database to which Elba connects. This database is interpreted as a many-valued context, hence every many-valued attribute has to be scaled, according to some rules, called conceptual scaling. The process of scaling once finished (for all, or just for some attributes of interest), a conceptual schema is provided. Siena just consideres the formal context output of conceptual scaling for further actions. The conceptual schema is then opened by ToscanaJ, which is the visualisation tool of this suite.

In our approach, we have used ToscanaJ [1] to build the conceptual hierarchis and then to browse the formed conceptual patterns over the data set. Elba [7] was used to build the scales over the database containing the PULSE accesses and to export schemas to be explored with Toscana. Browsing with Toscana offers the opportunity to use the available diagrams according to specific needs; it is possible to aggregate the diagrams in order to obtain results such as proof of the existence of patterns in attributes correlation. Different scenarios can be formed using only a small subset of the diagrams.

3. Formal Concept Analysis on web usage data

The web site used for collecting the usage/access data is an e-learning portal called PULSE [2]. The analysis was done only on the data collected from the two months of the last academic semester (i.e. April and May of 2012).

We have started our analysis by determining who is using PULSE and in what proportion. As depicted in Figure 2, students use this portal 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 appear under the label *'no login'* and take place just before the login phase when no login name can be recorded. Those 12142 accesses are made in the pre-login phase by both students and teacher (as all students and the teacher accessed the login page), but also by other individuals which did not have a PULSE account or landed on the PULSE login page accidentally.

The diagram can be also configured to show the number of accesses instead of login distributions. Thus, we can also determine the exact number of accesses as described above.

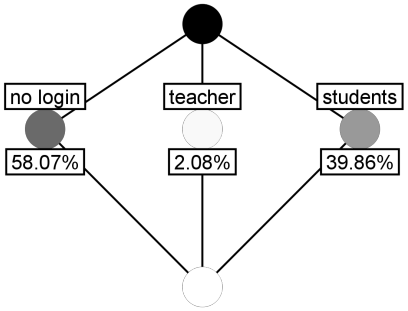

FIGURE 2. Who is using PULSE? Login distribution.

We continued our investigation by determining which students are using PULSE. Therefore, we generated the diagram from Figure 3 which divides student accesses by their accessed year of study.

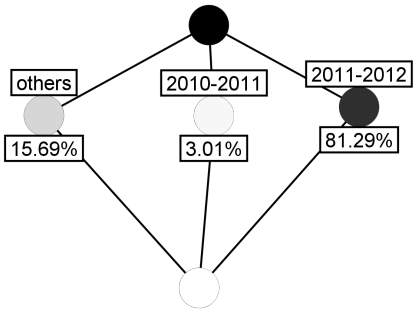

FIGURE 3. Which students? Years of study.

The results show that students revisit PULSE for checking the information taught in previous years. There were 6775 student accesses for currently taught subjects, 215 student accesses for subjects taught last, and 1308 student accesses for subjects taught in other previous years. We mention that the same subject on PULSE has a different content on different years of study as the teaching data is updated yearly, and each year a new set of students would have accounts created for the specific subjects. One student may have a single account but he/she can navigate through the different subject and years of study on which he/she has access on PULSE.

Figure 4 shows the access distribution of PULSE students on different subjects. As in the recorded semester (e.g. from February to June 2012), the subject studied was SO1 (i.e., Operating Systems), the percentage of almost 85% (=7015 accesses) is justifiable. What is interesting to observe here is that

the other subjects, 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 15.73%=1311 accesses were generated by visiting the pages before the subject selection phase.

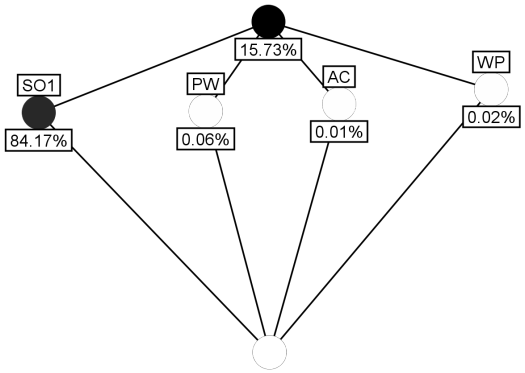

FIGURE 4. Teaching subjects for students

Over the considered period of time, the teacher accessed only the current subject as depicted in Figure 5. The 5.07% accesses are the ones recorded in each visit prior to the subject selection.

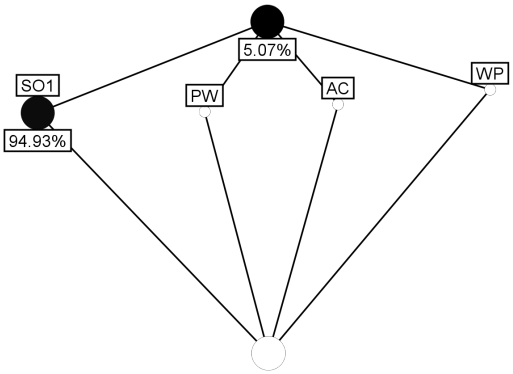

FIGURE 5. Teaching subjects for teacher

As a next step we wanted to check how well PULSE performs for the task that it was designed for. We wanted to see if students access the information provided. Figure 6 shows these results on different levels of importance. The distributions presented here confirmed our expectations. PULSE was designed as support instrument for laboratories and lectures. These classes have

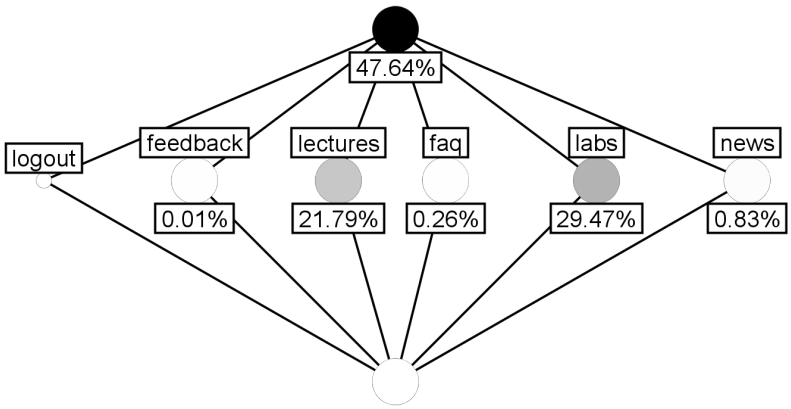

FIGURE 6. student PULSE accesses by classes

obtained the highest score. The 47.64% accesses were made on the PULSE *'homepage'*. This page loads just after the login phase and contains general information for students, such as:

- the name of the authenticated person;
- 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. The other facilities offered by PULSE are also presented in Figure 6. Students accessed 69 times the *'news'* page which contains all notifications/announcements made by the teacher for that specific subject. The last announcement is always posted also on the *'homepage'*. The *'faq'* (Frequently Asked Questions) page was accessed 22 times, while only one student access was made to send a *feedback*. A very insightful result is that students do not use the *logout* button.

ToscanaJ allows aggregating these diagrams and navigating from one diagram to another. For instance, if we select the node *'lectures'* from Figure 6, we can zoom into this node, in order to evaluate access distributions for lectures. The result is displayed in Figure 7a. In the same way, zooming into node *'labs'*, the result of access distribution is displayed in Figure 7b. Hence, by combining different scales (the above mentioned diagrams), we can build several scenarios of browsing the knowledge content of our database.

This browsing is reversible, at any point we may return and choose another scenario, highlighting other connections between our data. These scenarios might be understood as changing the point of view of our analysis.

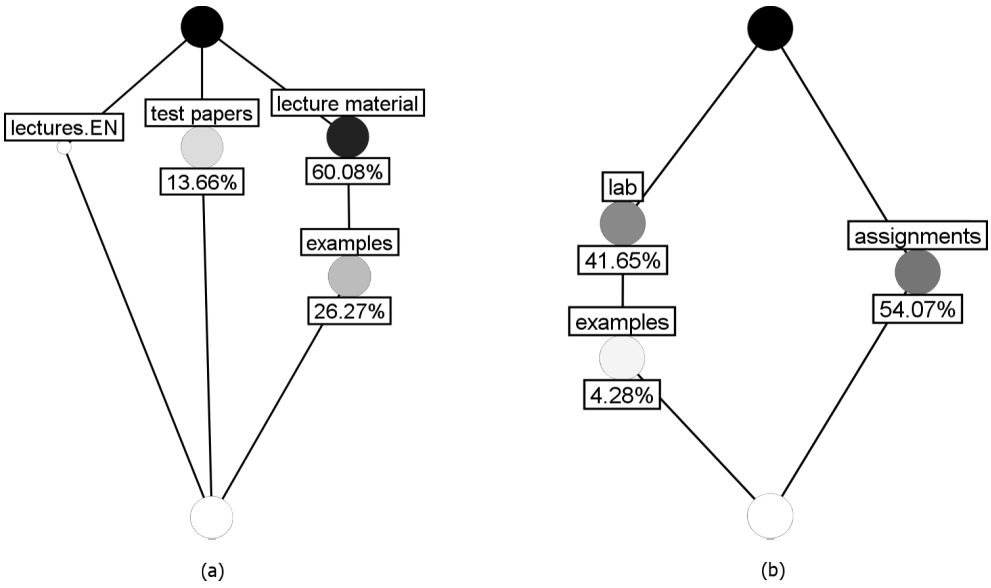

FIGURE 7. (a) access distribution for lecture related material, (b) access distribution for labs

As shown in Figure 7a, the most visited lecture related pages are those containing the theoretical support which consists *'lecture material'* and more *'examples'*, which means $60.08\% + 26.27\% = 86.35\%$. Then, there are the test papers during lectures and their results (including statistics and explanation how their marks will help the student) are presented under *'test papers'*. The detailed explanations on the solved test papers and statistics on the proportion in which the subjects of the test paper were solved by all students are placed together with the examples (i.e., under *'example'*. The lecture material is presented in Romanian. However, there are also English lectures. The

considered semester there were no English lectures, therefore there were no accesses for the node *'lectures.EN'*.

Figure 7b shows the detailed view on *'labs'*. Similar with lectures, here there is the technical support presenting the required concepts and *'examples'*. To better understand the concepts, students are given *'assignments'*.

PULSE facilities offered for teachers are presented in Figure 8.

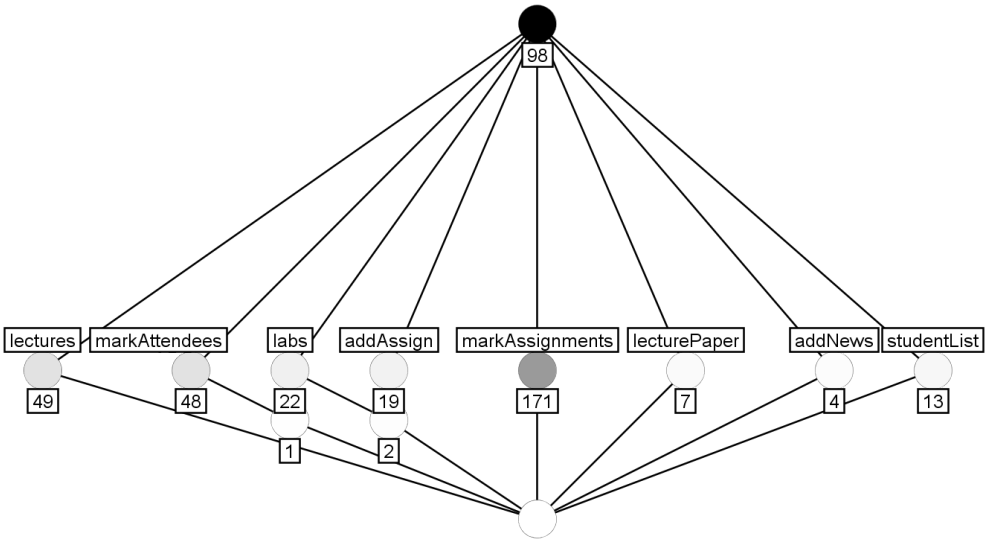

FIGURE 8. PULSE for teachers

Those facilities, in their most accessed order are:

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

Combining two or more conceptual hierarchy in one diagram is also possible using *nested-line diagrams*. These diagrams are obtained by partitioning

the attribute set in two or more subsets and then considering the direct product of the resulting conceptual sub-hierarchies. The original conceptual hierarchy is embedded in this direct product. This representation is particularly suitable if we would like to analyse the behaviour of our data with respect to a given collection of attributes. The following example depicted in Figure 9 shows a nested line diagram obtained by combining the months in the considered period and PULSE actors.

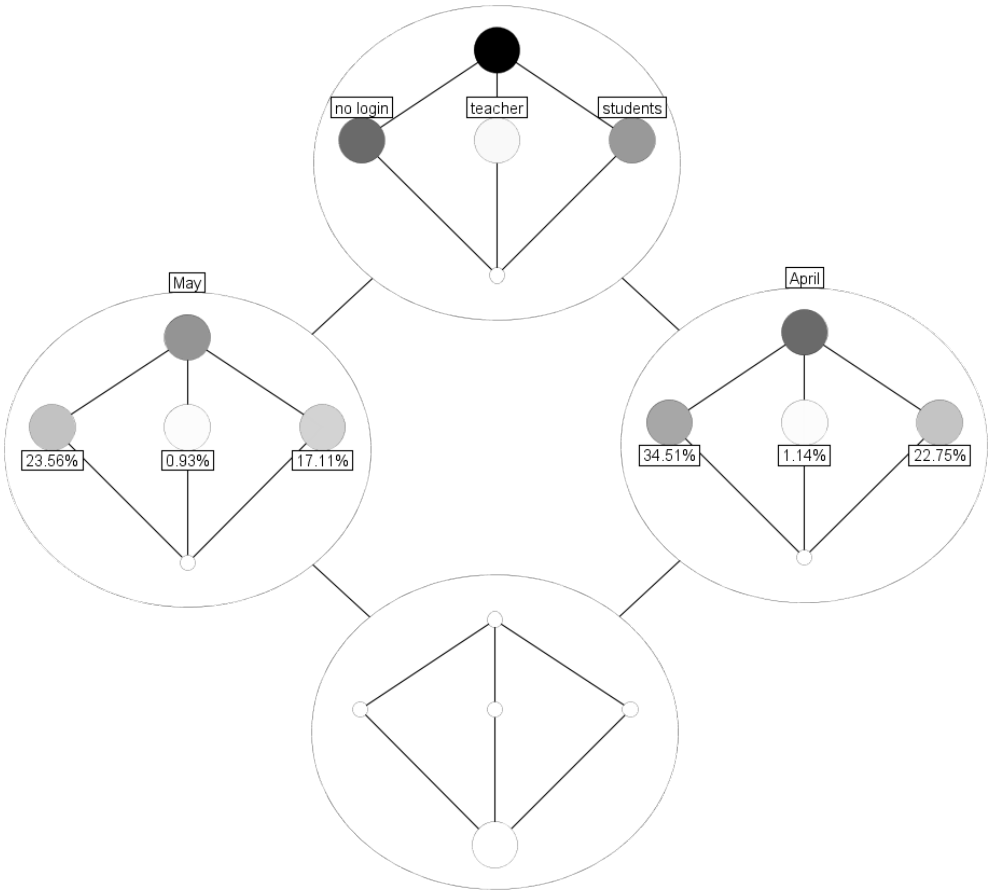

FIGURE 9. Nested diagram: months & login

4. CONCLUSIONS

The above considerations are reflecting a new approach into analysing web usage data. FCA allows more flexibility in analysing data by using predefined

concepts. Using the FCA diagrams one may navigate and interpret data from multiple different perspectives. By comparison, web analytics offers more limited perspective by using a set of metrics which can be interpreted based on their values.

We are now mainly interested in connections between our data and how the information gathered by the PULSE portal are linked and connected. Formal Concept Analysis seems to be a proper way to perform such analysis. Due to lack of space, we have not been able to describe several analysis scenarios, but more the main approach and analysis method, which have been, for the purpose of this paper of importance.

For a further research, we would like to investigate some of these relevant scenarios and to use conceptual logic, in form of attribute exploration and association rule mining might be helpful.

A possible development of this research might also be into the triadic setting of Triadic Formal Concept Analysis, but this will be presented in our future papers.

References

[1] P. Becker, J. Hereth, and G. Stumme, *Toscanaj - an open source tool for qualitative data analysis*, (2002).

[2] S. Dragos, *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.

[3] B. Ganter and R. Wille, *Formal concept analysis: mathematical foundations*, Springer Verlag, 1999.

[4] R. Wille, *Conceptual landscapes of knowledge: a pragmatic paradigm for knowledge processing*, in International Symposium on Knowledge Representation, Use, and Storage Efficiency, G. Mineau and A. Fall, eds., Vancouver, 1997, Simon Fraser University, pp. 2–13.

[5] ———, *Begriffliche wissensverarbeitung: Theorie und praxis*, Informatik Spektrum, (2000).

[6] ———, *Methods of conceptual knowledge processing*, in the 4th International Conference ICFCA, Springer Verlag, 2006, pp. 1–29.

[7] B. Wormuth, *Elba user manual.* Published online, 2004. http://kvo.uow.edu.au/kvopapers/Elba_User_Manual.pdf.

UBB Cluj-Napoca
E-mail address: sanda@cs.ubbcluj.ro, csacarea@cs.ubbcluj.ro