Applying Triadic FCA in Studying Web Usage Behaviors

Sanda Dragoş, Diana Haliţă, Christian Săcărea, and Diana Troancă

Babeş-Bolyai University, Department of Computer Science, Cluj-Napoca, Romania
sanda@cs.ubbcluj.ro, diana.halita@ubbcluj.ro, csacarea@math.ubbcluj.ro,
dianat@cs.ubbcluj.ro

Abstract. Formal Concept Analysis (FCA) is well known for its features addressing Knowledge Processing and Knowledge Representation as well as offering a reasoning support for understanding the structure of large collections of information and knowledge. This paper aims to introduce a triadic approach to the study of web usage behavior. User dynamics is captured in logs, containing a large variety of data. These logs are then studied using Triadic FCA, the knowledge content being expressed as a collection of triconcepts. Temporal aspects of web usage behavior are considered as conditions in tricontexts, being then expressed as modi in triconcepts. The gained knowledge is then visualized using CIRCOS, a software package for visualizing data and information in a circular layout. This circular layout emphasizes patterns of user dynamics.

1 Introduction

Investigating knowledge structures has a long tradition. In this paper, we propose an approach based on the Conceptual Knowledge Processing paradigm [1]. We use the idea of conceptual landscapes in order to highlight the visual part of organizing knowledge in a format which supports browsing and queries but also critical discourse. The implementation of such a system is thought to be a valuable help for the human expert, organizing knowledge in a way which supports human thinking and decision making.

Conceptual Knowledge Processing is an approach that underlies the constitutive role of thinking, arguing and communicating human being in dealing with knowledge and its processing. The term processing also underlines the fact that gaining or approximating knowledge is a process which should always be conceptual in the above sense. The methods of Conceptual Knowledge Processing have been introduced and discussed by R. Wille in [2], based on the pragmatic philosophy of Ch. S. Peirce, continued by K.-O. Apel and J. Habermas.

The mathematical theory underlying the methods of Conceptual Knowledge Processing is Formal Concept Analysis (FCA), providing a powerful mathematical tool to understand and investigate knowledge, based on a set-theoretical semantics, developing methods for representation, acquisition, and retrieval of knowledge. FCA provides a formalization of the classical understanding of a concept. Knowledge is organized in conceptual structures which are then graphically represented. These graphical representations are forming the basis for further investigation and reasoning.

In this paper, we apply Formal Concept Analysis to investigate web usage behavior. This study is conducted within a previously built conceptual information system (see Section 3). Herefrom, we select triadic data and compute a set of triadic concepts. These triadic concepts contain all relevant information related to knowledge structures encoded in the selected data set. We also use some derivation operators to process data for our web usage behavior study (see Section 4). In the last part of this paper, we focus on emphasizing patterns of user dynamics and their temporal behaviour using a circular visualization tool.

2 Prerequisites: Triadic Formal Concept Analysis

In the following, we briefly recall some definitions. For more, please refer to [3] and [4]. The fundamental data structure triadic FCA uses is *a tricontext*, which exploits the fact that data might be represented in 3D tables of objects, attributes, and conditions. Hence, a *tricontext* is a quadruple $\mathbb{K} := (K_1, K_2, K_3, Y)$ where K_1, K_2 and K_3 are sets, and Y is a ternary relation between them, i. e., $Y \subseteq K_1 \times K_2 \times K_3$. The elements of K_1, K_2 and K_3 are called (formal) objects, attributes, and conditions, respectively. An element $(g, m, b) \in Y$ is read *object g has attribute m under condition b*. Every tricontext has an underlying conceptual structure reflecting the knowledge which is encoded in the triadic data set. This conceptual structure is described by the so-called triconcepts. In order to mine them, derivation operators are defined ([4]). Every triconcept is a maximal triple (A, B, C), where $A \subseteq K_1, B \subseteq K_2, C \subseteq K_3$ having the property that for every $a \in A, b \in B, c \in C$, we have $(a, b, c) \in Y$.

3 Web Usage Mining and Previous Work

A large amount of collateral information about web usage information is stored in databases or web server logs. Statistics and/or data mining techniques are used to discover and extract useful information from these logs [5].

Web analytics tools are based on some web analytics metrics. They prove to be a proper method to give a rough insight about the analysed web site, especially if it is a commercial site. However, the purpose of an e-commerce site is to sell products, while the purpose of an e-learning site is to offer information. Therefore, a visit on an educational site does not apply to the heuristics used by most analytics instruments [6].

Web usage mining focuses mainly on research, like pattern analysis, system improvement, business intelligence, and usage profiles [7, 8]. The process of web usage mining undergoes three phases: preprocessing, pattern discovery, and pattern analysis. At the preprocessing phase data is cleaned, the users and the sessions are identified and the data is prepared for pattern discovery. Such usage analysis constitutes an important feedback for website optimization, but they are also used for web personalization [9, 10] and predictions [11].

The web site used for collecting the usage/access data is an e-learning portal called PULSE [12]. The web usage data collected from PULSE was already processed by using Formal Concept Analysis [13], where a detailed description of using ToscanaJ to build a conceptual information system for a previous version of PULSE is given.

The analysis is performed on the data collected from the second semester of the academic year 2012-2013 (i.e., from the beginning of February 2013 to the end of July 2013). A log system records all PULSE accesses into a MySQL database. For the analysed time interval there were 40768 PULSE accesses. The data fields from the collected information used in the current investigation are Full request-URI, Referrer URL, Login ID and Timestamp.

The data to be analysed contains 751 distinct request-URIs (i.e., access files), 471 distinct referrers, 130 distinct login IDs and 25798 distinct timestamps. Using the same methodology as in [13], a ToscanaJ conceptual information system has been built over the PULSE log files. Each data field has been scaled and a conceptual scale has been created. The datasets are considered many-valued contexts, the semantics of attributes being expressed by *conceptual scales.*

For this research, we are interested in investigating temporal patterns of web usage behavior within PULSE. Hence, we will restrict our focus only to the access file classes, the referrer classes and the timestamps of the system. This is a natural triadic structure whereof we can extract user dynamics related knowledge structures in form of triadic concepts.

3.1 Access File Classes:

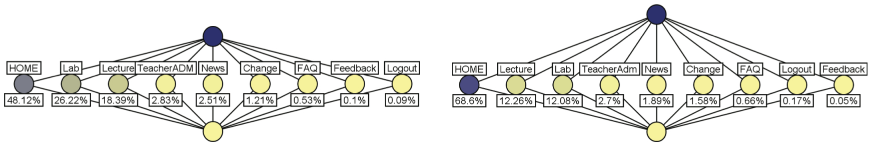
The request-URI represents the address of the accessed webpage along with all query information used for that actual request. Although we value the information contained by this field, the granularity of the accessed web pages is too fine for our intent (i.e., there are 751 distinct access file entries in the database). Thus, the accessed webpages have been divided into 9 classes (see Figure 1(a)).

PULSE portal was intended to be used mainly during laboratory sessions for students to access appointed assignments with the related theoretical support or to consult lecture related content. Each user enters PULSE through the HOME page which contains general information such as: lab attendances, marks, evaluation remarks and current announcements. All webpages related to the content described above are grouped into the three main classes named: **Lab**, **Lecture** and **HOME** according with their purpose.

The other 6 classes represent administrative utilities for the teacher grouped into the **TeacherAdm** class and informative sections for the students, such as **FAQ** (i.e., frequently asked questions), **Feedback**, **News**, **Logout** or navigation through the content of the course from previous years of study (**Change**).

3.2 Referrer Classes:

Referrer URLs represent the webpage/site from which the current webpage/access file was accessed. The referrers which are outside of PULSE are not used in this current research. The referrers which represent PULSE webpages fall into the same classes as access files do. The access file classes and referrer classes have been scaled nominally and visualized with ToscanaJ in Figure 1(b).



(a) Access File classes

(b) Referer classes - 95.7% are from inside PULSE

Fig. 1: Nominal scales of Access File and Referrer classes

4 Web Usage Mining with Triadic FCA

The extension Toscana2Trias allows the selection of triadic data starting from a given set of scales, if the data has been preprocessed with ToscanaJ. From the conceptual schema file, we have selected the scales presented above and obtained a triadic data set using the pairs Referrer class-Access File class (R_class-AF_class) as attribute set, timestamps as conditions and students Login as object set. Then, we have generated all triconcepts using the Trias algorithm [14].

The problem of visualizing triadic data has not been yet satisfactory solved. Triadic conceptual structures have been visualized for instance using trilattices or graphs. Circos as a visualizing tool has been developed to investigate structural patterns arising in bioinformatics. In this section, we present a proof of concepts in order to show possible applications of using Circos to visualize triadic content.

4.1 Interpreting triadic FCA results with Circos

Circos is a software package for visualizing data and information in a circular layout. This circular layout emphasizes patterns in the dataset, showing connections between represented data [15].

The input data for Circos is obtained from the tricontext using a derivation operator. We implemented an algorithm that analyzes the XML output of Trias and transforms it into a valid input for Circos.

The XML file that results as an output from Trias contains all triconcepts which can be derived from the tricontext defined over the data set using Toscana2Trias. Each of them is defined by an extent, an intent and a modus. The valid input data set for Circos is a bidimensional table $R \times C$, with numerical values, hence we have to derive these tables from the tricontext.

Starting from the tricontext (G, M, B, Y), we first build a dyadic projection $\mathbb{K}_{32} := (G, (B, M), I)$, where $(g, (b, m)) \in I \Leftrightarrow (g, m, b) \in Y$. Then, for each pair

(b, m) we compute the corresponding attribute concept $\mu_{\mathbb{K}_{32}}$ and determine the cardinality of its extent $(b, m)'$.

The set of column indicators, denoted C, is the set obtained by projecting the ternary incidence relation Y on M, $\mathrm{pr}_M(Y) := \{m \in M \mid \exists (g, b) \in G \times C.\ (g, m, b) \in Y\}$. Similarly, the set of row indicators, denoted R, is the set obtaining by projecting Y on the set of conditions B.

The algorithm we have implemented builds a table having these sets as column and row indicators and calculates the numerical values of the table as follows. For each pair $(c, r) \in C \times R$, the cardinality of the extent $(c, r)'$ in $\mathbb{K}_{32}$ is computed directly from the XML output of Trias. This cardinality represents the numerical value of the cell corresponding to the column c and the row r.

As a final step, we visualize our data by running Circos and obtain an output in png or svg format. Figure 2 presents web usage of the student group "ar" on the 10th week as described in Section 5. Each ribbon corresponds to a pair (Referrer class, Access File class). Because the set of referrer classes and the set of access file classes have elements in common, the sets C and R are not disjoint.

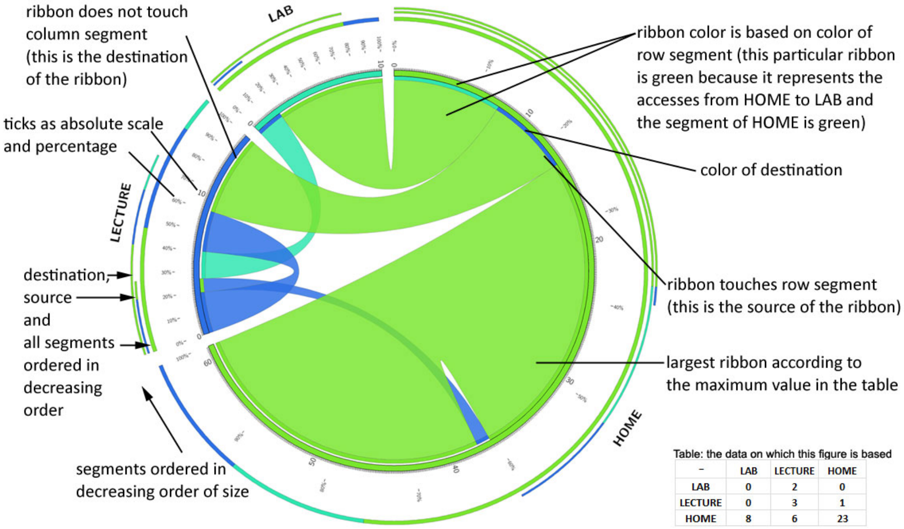


Table: the data on which this figure is based

–	LAB	LECTURE	HOME
LAB	0	2	0
LECTURE	0	3	1
HOME	8	6	23

Fig. 2: Results for the "ar" students on the 10th week of school

5 Test results

The data used was gathered during an entire semester and because of the high volume of data, one single circular representation did not reveal any useful patterns. Therefore, we reduced the volume of data and aggregated the object set (i.e., individual student login) into the set of student groups, as distributed according to their curricula. We continued our tests treating each group of students separately, considering R_classes as objects set, AF_classes as attribute set and timestamps as conditions. In order to investigate the temporal behaviour of students during one semester, we analysed the data on time intervals.

The first time granule we have considered was approximately one third of a semester (beginning, middle and the end). Such time intervals do not provide any significant patterns, and therefore we fine-tuned the time granule to a week.

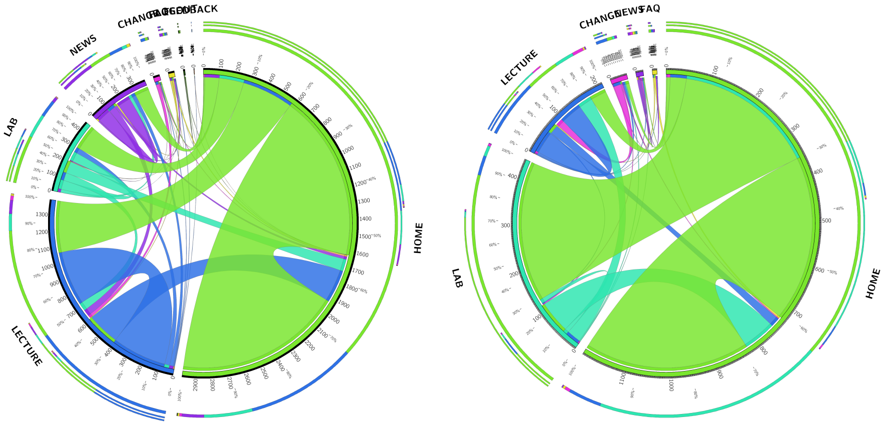
The PULSE portal recorded data on two courses for the semester we have considered: Operating Systems (SO1), which is a compulsory course and Web Design Optimisation (WDO), which is an elective course. Two student groups enrolled in the SO1 course, denoted ”ar” and ”ri”. For the WDO course, students from five different groups enrolled. WDO being an elective course, some of the student groups were poorly represented, hence we studied the behaviour of two of these student groups, denoted ”ei” and ”ie”.

The entire set of results are posted at `http://www.cs.ubbcluj.ro/~fca/ksemtests-2014/`. We observed from these results that there are three types of behavior, which we named: relaxed, intense and normal.

The relaxed behaviour occurs mainly during holiday (e.g., the 10th week). The pattern for this type of behavior is depicted in Figure 2 and can be distinguished by the fewer accesses and the reduced number of Access File classes visited (e.g., usually only the main classes). The navigational patterns observed during this week were really simple. For instance for the ”ar” group of students went from HOME to either LAB or LECTURE; from LAB they went to LECTURE, and from LECTURE they went back to HOME. For the elective course (i.e., WDO) the results show more relaxed patterns due to the fact the this type of course implies personal research. Therefore, the teaching material provided is less visited than in the case of the compulsory course (i.e., SO1). This type of behavior occurs also after final exams or between the final exam and the re-examination: 18th and 20th week for group ”ar”, 18th and 19th week for group ”ri” and after the 14th week for groups ”ei” and ”ie”.

The intense behavior occurs during examination periods. The pattern depicted in Figure 3(a) shows an increase number of accesses. The pattern can be observed even in the weeks before the examination, its peek occurring during the week of the exam. It is the case of the 17th week for the ”ar” and ”ri” groups for the final exam, and the week 19th for ”ar” group and week 20th for the ”ri” group for re-examination. For the elective course (i.e., groups ”ie” and ”ei”) there are three evaluation dates during the weeks 7th, 9th and 13th.

The normal behavior occurs during the semester when there is no examination period or holiday. The pattern for this type of behaviour, as show in Figure 3(b), is that almost all Access File classes are visited. The three main classes contain the most visited pages. The next most visited class is News. These results are to be expected as PULSE is mainly intented to provide support for laboratory, lectures and to post news.



(a) 17th week - intense

(b) 5th week - normal

Fig. 3: Comparative behaviors on "ar" group: intense versus normal

Although in Figure 3 the two behaviors can appear similar, there are some important differences. The main difference is the fact that during the intense period the webpages from the Lecture class are more visited as the students prepare for the examination, while in the normal period the webpages in the Lab class are most visited as students solve their lab assignments. Another difference is the fact that even if HOME is the most visited class, during the intense period it looks like it represents merely a connection to the other PULSE facilities (i.e., Lecture, Lab, News). The number of accesses is another difference as there are much more accesses during the intense periods.

The triadic conceptual landscape as computed by TRIAS provides a large amount of information that is suitable for a large variety of interpretation/visualization. Histograms are also provided at `http://www.cs.ubbcluj.ro/~fca/ksemtests-2014/`. This representation however, presents only the quantitative aspect of the navigation meaning the number of accesses. The circular visualization presented so far provides a more qualitative view on the navigational pattern, comprising more details about how and where students navigate.

6 Conclusion and Future Work

The research conducted so far and the previous related work, shows how triadic conceptual landscapes can be used for web usage mining and representation of user dynamics patterns. The main advantage of using conceptual landscape versus different interrogations rely on the completeness of the information clustered in a concept, or determined by a derivation, respectively. Circular visualization tools can be applied to any quantitative data, the qualitative interpretation comes from the conceptual preprocessing.

For further research, we will apply the methods of Temporal Concept Analysis and develop them for the triadic setting with the aim to describe user trails, life tracks and bundles of trails and tracks.

References

1. Wille, R.: Conceptual Landscapes of Knowledge: a Pragmatic Paradigm for Knowledge Processing, In: Gaul, W.; Locarek-Junge H. (Eds.): Classification in the Information Age, Proceedings of the 22nd Annual Gfki Conference, Dresden, March 4-6, 1998, pp. 344–356.
2. Wille R.: Methods of Conceptual Knowledge Processing. In: Missaoui R., Schmidt, J. (Eds.) 4th International Conference ICFCA 2006, Dresden, Germany, LNAI, vol. 3874, pp. 1–29, Springer Heidelberg (2006).
3. Ganter B., Wille R.: Formal Concept Analysis. Mathematical Foundations. Springer, Berlin-Heidelberg-New York (1999).
4. Lehmann F., Wille R.: A Triadic Approach to Formal Concept Analysis, In: Ellis, G., Levinson R., Rich. W., Sowa J. (eds.) Conceptual Structures: Applications, Implementation and Theory, LNCS, vol. 954, pp. 32 – 43, Springer, Heidelberg (1995).
5. Kosala R., Blockeel H.: Web Mining Research: A survey, In: ACM SIGKKD Explorations Newsletter, vol. 2, pp. 1–15, ACM, New York (2000).
6. Dragoş S.: Why Google Analytics Can Not Be Used For Educational Web Content. In: Abraham A. et all. (Eds.) 7th International Conference on Next Generation Web Services Practices (NWeSP), pp. 113–118, IEEE (2011).
7. Spiliopoulou M., Faulstich L. C.: Wum: a Tool for Web Utilization Analysis. In: The World Wide Web and Databases, LNCS, vol. 1590, pp. 184–203, Springer, Berlin Heidelberg (1999).
8. Srivastava J., Cooley R., Deshpande M., Pang-Ning T.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. In: ACM SIGKDD Explorations Newsletter, vol. 1, Issue 2, pp. 12–23, ACM, New York (2000).
9. Eirinaki M., Vazirgiannis M.: Web Mining for Web Personalization. In: ACM Transactions on Internet Technology (TOIT), vol. 3, Issue 1, pp. 1–27, ACM, New York (2003).
10. Romero C., Ventura S., Zafra A., Bra P. D.: Applying Web Usage Mining for Personalizing Hyperlinks in Web-Based Adaptive Educational Systems. In: Computers & Education, 53, 828–840 (2009).
11. Romero C., Espejo P. G. , Zafra A., Romero J. R., Ventura S.: Web Usage Mining for Predicting Final Marks of Students That Use Moodle Courses. In: Computer Applications in Engineering Education, 21,135–146 (2013).
12. Dragoş S.: PULSE Extended. In:The Fourth International Conference on Internet and Web Applications and Services, Venice/Mestre, Italy, May 2009, pp. 510–515, IEEE Computer Society (2009).
13. Dragoş S., Săcărea C.: Analysing the Usage of Pulse Portal with Formal Concept Analysis. In: Studia Universitatis Babes-Bolyai Series Informatica, vol. LVII , pp. 65–75 (2012).
14. Jaeschke R., Hotho A., Schmitz C., Ganter B., Stumme G.: Trias - An Algorithm for Mining Iceberg Trilattices. In: Proceedings of the IEEE International Conference on Data Mining, pp. 907-911, Hong Kong, IEEE Computer Society (2006).
15. Circos, a circular visualization tool,`www.circos.ca`