PROCESS-ORIENTED ANALYSIS OF WEB-BASED DATA

František BABIČ, Ján PARALIČ

Technical University of Košice
Faculty of Electrical Engineering and Informatics
Department of Cybernetics and Artificial Intelligence
Letná 9/B, 042 01 Košice, Slovakia
{Jan.Paralic,Frantisek.Babic}@tuke.sk

Abstract. This paper presents some proposed approaches to analyse web-based data from the process analysis perspective. It summarizes our experiences and results obtained during implementation of a couple of research projects, mainly FP6 IST project called KP-Lab and national Slovak project called TraDice. Process analysis is represented by methods for processing collected historical data and extraction of potentially useful information or knowledge from these datasets. The aim is to create a historical projection of performed collaborative activities with visualization of relations between all involved persons, used shared objects of interests or created results. Complex understanding of this "live" environment provides important source of knowledge for its evaluation or further improvement based on e.g. identified best or worst practices.

1. Introduction

Collaborative activities mediated by various web-based environments are driven by various shard objects and models of user behaviour. They provide a very rich source of information that can be extracted and analysed based on defined analytical goals, e.g. identification of network of collaboration behind successful goals achievement; extraction of hidden patterns that lead to particular good or bad practice; creation of possible useful and helpful recommendations based on previous experience extracted from historical data; visualization of the whole performed processes through timeline to provide simple understanding for less experienced users; etc.

We provide some simple examples that show our motivation and starting point for design and implementation of described analytical approaches. The first presents a typical working or learning session mediated by suitable collaborative environment. This session is oriented around relevant shared objects e.g. documents, video or audio files, programming codes, demos, etc. Shared objects of interests can be divided into inputs and results

of activities performed by users. Users have a different theoretical background, knowledge or method of working, but together they will work as a team to support effective collaborative work or learning. If we are able to monitor and store all performed actions with relevant shared objects, we will create an interesting historical dataset for analytical purposes. In this case, we will be able to identify key inputs for decisions, the most active or less active participants, crucial persons in network of collaboration and last but not least we will be able to create recommendations for similar future sessions based on identified best or worst practices.

The second example, describes travelling of users in virtual space. Based on collected data representing all performed actions within virtual library, we will be able to extract a model of user behaviour within this type of virtual space. This model will include user's search goals with related trajectories; shared objects of the second level that don't represent a goal of primary search, but will be open based on their relations to the primary trajectory and lists of professional interests or skills. Models can be further clustered based on identification of similar characters and we will be able to create a typical models of behaviour for different domains of interests, e.g. if a typical researcher from domain of process mining will start the search, we can provide him some starting materials to simplify his travel through very complex virtual space, or recommend relevant next readings in particular user's context.

1.1 Related Work

During past several years devoted to our research described in this article we have identified many interesting approaches and tools that are oriented to the similar goals. In this section we will present only some selected research works with relevant results. Customized web usage mining is described in [13] to support and enhance three main stages as data gathering; preprocessing and pattern discovery with possibility to express specific constraints in order to e.g. reduce the search space or to make the final pattern simpler and more user friendly.

Interaction analysis is represented by PAnDit "Pattern Analysis and Discovery Tool" that works with user activities stored in custom log format in combination with patterns defined as rules in Prolog [9]. Users can search for occurrences of various groups of events, or create a nontrivial filter to select interesting events. All rules are stored and applied across different studies and contexts. This tool works with structured format of logs (CoLoForm) agreed within European Kaleidoscope network of excellence that offers some other interesting analytical approaches as Synergo interaction analysis tool or Activity Lens [5, 12]. Synergo provides three main data views: quantitative, qualitative and playback of activity that can be used to create a complex overview or historical reconstruction of all performed activities. Other automatic possibilities to analyze performed activities in groupware or CSCL systems are described in [6, 7, 8].

Learning scenario represents an important part of collaborative activities and its structure depends on many factors such as conditions, objectives, experiences, expectations, etc. Some of these factors can be extracted from previous instances of the related activities for the future customization and improvement. And this is the aim of an intelli-

gent authoring tool called CHOCOLATO [10]. Quite different but interesting approach represents a fuzzy expert system for evaluation of virtual collaboration described in [11]. This approach is based on combination of collected logs from virtual environment and evaluation through predefined hierarchical fuzzy rules representing expert knowledge and experience.

2. Description of Proposed Approaches

This section will briefly present proposed approaches for manipulation and processing of collected historical data in the specified log format. This log of events provides a basic structure designed in order to ensure a generic enough format for storing of the data. The initial structure consists of twelve parameters, e.g. ID, type of action, type of shared object, type of user, user ID, object ID, time stamp etc. Detailed information about logging features and log format can be found in [1, 3].

2.1 Historical projection

Historical projection provides an interesting approach based on visual analytics paradigm to visualize performed actions in chronological order with all relevant shared objects and participants that dealt with them. The aim is to visualize sequences of performed steps in virtual space in order to achieve specific goals. It is possible to create several parallel timelines according to number of investigated participants. Different types of events are distinguished by graphical icons, e.g. star is equal to "modification" or circle to "opening". Displayed timeline is interactive; users have possibilities to filter visualized events or to highlight selected events based on specified conditions. Detailed information and graphical demonstration are presented in [1, 3, 4]. Important part of this approach is a possibility to define a pattern as well formalized projection of interesting practices.

2.2 Patterns

Patterns can be understood as formalization of captured hidden knowledge in the activity specified by user based on his experiences and background knowledge. The pattern is represented as a sequence of pattern elements, where each pattern element represents one generalized event. In the pattern element, which is essentially a list of key-value pairs, user specifies which parts of the generalized event are important and which should be generalized. User can specify the element based on any subset of the event's properties, including custom ones. Defined pattern is matched with events stored in log repository to generate a search tree in which obtained results are represented as leaf nodes of the tree. More details are available in [2, 3, 4].

2.3 Summary statistics

Another visual analytics approach in this domain are summary statistics provide a basic overview of collected data based on combination of user queries and visualization of their results by means of different graphs. These queries are constructed from parameters of log format in order to meet user requirements. Defined queries are executed over the database of logs and obtained results are visualized in form of selected type of graph. User can in such a way visualise basic statistics characterizing performed activities in virtual space,

e.g. how many shared objects were modified by this user, how many tasks were assigned to this user, which user is the most active during the session, etc. Some particular examples are available in [2, 4].

3. Conclusions

In this paper we summarized the results of our continuous research oriented to different approaches and proposed methods for analysis of "process" data. The common goal is to provide simple and understandable services with potential to be applied to different application domains as web, virtual learning environments, virtual libraries or e.g. environment for software testing. Our future work will deal with investigation of opportunities for integration of the proposed approaches with techniques for manipulation and processing of big data.

Acknowledgement: This work was partially supported by Slovak Research and Development Agency under the Contract No. APVV-0208-10; Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic under Grant No. 1/1147/12; Research & Development Operational Program funded by the ERDF through implementation of the project called Development of the Center of Information and Communication Technologies for Knowledge Systems (ITMS: 26220120030); and European Commission DG INFSO under the IST programme, Contract No. 27490. The KP-lab integrated project was sponsored under the 6th EU Framework Programme for Research and Development. The authors are solely responsible for the content of this article. It does not represent the opinion of the KP-lab consortium or the European community, and the European community is not responsible for any use that might be made of data appearing therein.

References

to other papers publishing the results that are summarized here

[1] Babič, F., Wagner, J., Paralič, J.: The use of event logs for collaborative practices reflection. In: *International Journal of Intelligent Information and Database Systems*, (2012), vol. 6, no. 5, pp. 421-435.

[2] Babič, F., Wagner, J., Paralič, J.: Investigation of performed user activities in overall context with IT analytical framework. In: *Lecture Notes in Business Information Processing: Business Information Systems*, Heidelberg: Springer-Verlag, (2012), vol. 117, pp. 284-295.

[3] Paralič, J., Richter, Ch., Babič, F., Wagner, J., Raček, M.: Mirroring of knowledge practices based on user-defined patterns. In: *The Journal of Universal Computer Science*, (2011), vol. 17, no. 10, pp. 1474-1491.

[4] Paralič, J., Babič, F., Paralič, M.: Process – driven approaches to knowledge transformation. In: *Acta Polytechnica Hungarica*, (2013), vol. 10, no. 5, pp. 125 – 143.

Other references

[5] Avouris, N., Kahrimanis, G., Fiotakis, G., Voyiatzaki, E. and Margaritis, M.: Tools supporting collaborative learning and interaction analysis: Synergo and ActivityLens. *Symposium on Designing, Re-using and Exchanging Online Learner Corpora*, (2007), Grenoble.

[6] Bratitsis, T. et al.: Supporting members of a learning community using interaction analysis tools: the example of the kaleidoscope NoE scientific network. In: *Eighth IEEE International Conference on Advanced Learning Technologies*, (2008), pp. 809–813.

[7] Duque, R. et al.: Modelling and capturing users' actions in CSCL systems for analysis purposes. In: *International Journal of Emerging Technologies in Learning* (iJET), (2009), vol. 4, pp. 53 – 59.

[8] Duque, R., Bravo, C., Ortega, M.: An ontological approach to automating collaboration and interaction analysis in groupware systems. In: *Knowledge-Based Systems*, (2013), vol. 37 pp. 211 – 229.

[9] Harrer, A., Lingnau, A., Bientzle, M.: Interaction Analysis with dedicated logfile analysis tools – a comparative case using the PAnDit tool versus manual inspection. In: *Ninth IEEE International Conference on Advanced Learning Technologies*, IEEE, (2009), pp. 405—407.

[10] Isotani, S. et al.: A Semantic Web-based authoring tool to facilitate the planning of collaborative learning scenarios compliant with learning theories. In: *Computers & Education*, (2013), vol. 63, , pp. 267 – 284.

[11] Strigunaite, S., Kriksciuniene, D.: Fuzzy expert system for virtual team collaboration and work evaluation. In: *Lecture Notes in Business Information Processing*, (2012), vol. 127, pp. 37 – 43.

[12] Voyiatzaki, E., Margaritis, M. and Avouris, N.: Collaborative interaction analysis: the teacher's perspective. In: *Proc. ICALT 2006 – Proc. 6th IEEE International Conference on Advanced Learning Technologies*, Kerkrade, Netherlands, pp. 345 – 349.

[13] Zaïane, O.R. and Luo, J.: Towards evaluating learners' behaviour in a web-based distance learning environment. In: *2nd IEEE International Conference on Advanced Learning Technologies* (ICALT'01), Madison, USA, pp. 357 – 360.