QUANTITATIVE METHODS IN ECONOMICS
Vol. XII, No. 1, 2011, pp. 77–86

IDENTIFICATION OF WEB PLATFORMS USAGE PATTERNS WITH DYNAMIC TIME SERIES ANALYSIS METHODS

Jarosław Jankowski
Faculty of Computer Science and Information Technology
West Pomeranian University of Technology in Szczecin
jjankowski@wi.zut.edu.pl

Summary: The paper proposes a new approach to modelling online social systems users' behaviours based on dynamic time wrap algorithm integrated with online system's databases. The proposed method can be applied in the field of community platforms, virtual worlds and massively multiplayer online systems to capture quantitative characteristic of usage patterns.

Key words: social platforms, time series analysis, web users' behaviour

INTRODUCTION

Analyses of behaviour of Internet systems users plays a crucial role in decision-making processes and management of online platforms. This also relates to users of social services, in which analyses enable users to get to know the behaviour patterns and determine users' needs in a more efficient way. The research can be done in many scopes, and there are different approaches applied concerning data processing that are based on web mining algorithms or on social network analysis. In services focused on communities, the significant area that is a subject to exploration is the activeness of users and identification of behaviour patterns. The article herein presents new approach based on dynamic time wrap algorithm integrated with online system database that identifies two dimensions of temporal data analysis based on real and virtual time that enables the segmentation of users focused on time conditioning and trends identification. Research conducted in the real environment indicates numerous areas of applications and gives the basis for development of new methods and theoretical models. The proposed method allows the comparison of sequences in relation to the ideal pattern that represents the maximum possible number of logons or reference to the behaviour of other users.

MOTIVATION AND RELATED WORK

The development of social services, online games systems, and virtual worlds creates the demand of implementation of new methods for processing of data focused on those environments. The analyses aim to recognize different occurrences and trends and to acquire knowledge that can have a significant role in the development of online services. In this trend there is increased interest in, among others, methods of social network analysis, which applications in relation to Internet web systems is numerously present in literature concerning detection of social leaders and communities structures [Newman 2006][Newman, Girvan 2006], diffusion and marketing message [Marciniak, Budnarowska 2009][Acar, Polonsky 2008], recommendation message [Jung 2005][Gursel, Sen 2009][Che, Hu 2010] or dynamics of network structure development [Marsili, Slanina 2010], as well as in other fields. There are also developed solutions deriving from methods of knowledge discovery in databases, which with reference to Internet systems acquire a specialized form of Web mining algorithms, which are widely discussed in literature. They can be applied, among others, at the basis of agent systems [Gomes, Canuto 2006][Chau et al. 2003], in recommending systems [Kazienko 2009], electronic marketing [Chun-Ling et al. 2010], and in design of intelligent interfaces and adaptative portals. In most kind of appliances, the data has a temporal form, and time conditioning is taken into consideration. According to A.S. Dick and K. Basu, identification of consumers and increasing their loyalty constitutes a significant element of marketing actions [Dick, Basu 1994]. What is also important in creating loyalty towards advanced online services is the habit and limitations of introduction threshold where getting to know a new system is time-consuming. According to H. Tsai and H. Huang [Tsai, Huang 2007], it can discourage searching for alternatives. F. Wangenheim and T. Bayona state that behavioural aspects of loyalty translate into the purchase of a service and recognition of the superiority of an individual service over other subjects on the market [Wangenheim, Bayon 2004]. The loyalty of Internet system users constitutes a subject of separate research. H.P. Lu and S. Wang identify a dependency between users' satisfaction and loyalty in connection to Internet addiction [Lu, Wang 2008]. The development of technologies based on communities increases users' interactions and can be an element of strategies focused on loyalty building, e.g. in a form of weblogs in the scope of corporative systems, which was indicated by S.C. Herring and others [Herring et al. 2005]. Interpersonal engagement is an additional factor of loyalty building; bonds occur not only on the line of users - technology supplier, but also in the scope of technological platform, which was proved by C. Wagner and N. Bolloju [Wagner, Bolloju 2005]. Revealing hidden information and its time characteristic is a crucial element of data processing. Its specification both in interpretation and representation aspect requires the usage of new approaches that provide better possibilities of occurrence modelling and identification of its characteristics. Temporal data mining is one of the developing fields of methods for knowledge extraction from data bases focused on time characteristics. C.M. Antunes and A.L.Oliveira present a review of approaches and methods used in this field [Antunes, Oliveira 2001]. The methods of time representation and scope of appliance in different fields were identified. The review of methods presented in the study relates to temporary data bases and indicate development direction of methods focused on acquiring knowledge about hidden patterns and analysis of changeability of sets with temporal characteristics [Roddick, Myra 2002]. The unresolved problems of data exploration were also presented. The study [Hu et al. 2007] develops the aspect of Web data specification and the need to search for dedicated solutions that include local solutions, which in case of searching for generalized dependencies for all data set are not detected. The presented methods are focused on acquiring temporal, indirect, frequent patterns and their temporal, extended patterns in the area of identifying Web users with distinct interests.

As B. Hernandez-Ortega, J. Jimenez and others proved, there are significant differences in behaviour of potential, new or permanent consumers [Hernández-Ortega et al. 2008]. The identification of their behaviour enables adequate segmentation and offers targeting or adaptation of functionality to a given group of recipients. S.M. Beitzel and others presented the analysis focused on the temporal aspects and arrangement in time of data from logs from servers connected with search engines, and they proved that there is a possibility of temporal analyses in this environment [Beitzel et al. 2004]. Changeability of consumers' behaviour in time and evolution in time connected with different factors, as well as developing new technologies in time, is emphasized in the works of V. Venkatesh and M.G. Morris [Venkatesh et al. 2003], E. Karahanna and others [Karahanna 1999]. The concept of analysis of time series focused on analysis of Web search system logs was presented by Y. Zhang, B.J. Jansen and others [Zhang et al. 2009]. The analysis of time series enabled the detection of dynamic occurrences and consumers' behaviour patterns. The presented studies focus on the aspect of representation and temporal data analyses. Nevertheless, they do not include the specification of social services functioning and functioning of users in virtual reality, for which different structuring of time lapse can be determined than in real systems, and with a use of dimension analysis, the evaluation of loyalty level or characteristic of using individual service or platform can be done.

CONCEPTUAL FRAMEWORK

The article herein proposes a new approach towards data analysis and characteristics of web platforms usage, which relate to the users' activeness and engagement. Internet systems focused on communities very often function in a form of asynchronous virtual worlds where users create their identities and general activeness for their virtual beings. In systems of that type, the communication is done asynchronously, contrary to virtual worlds like Second Life, where data transmission is done in direct mode. Both in the first and the second case, the user's virtual representation functions in other reality in a sense of place (cyber space), so it can be assumed that the virtual representation functions also in another dimension of time. Lapse of time in a virtual system herein referred to as virtual time can be considered regardless of real time. For mapping and parameters' determination purposes, the two-dimensional model of time representation in bi-temporal scope was assumed, where occurrences are registered in a scope of real and virtual time. There is an activated snapshot of users' states and activities with assigned virtual world timestamp. The significant matter is to determine granulation of virtual time and its relation to real time. In the easiest presentation the unit of one virtual day lapse can directly reflect a unit of real time and can be identical with a real day. There can be different time separations introduced that will lead to an increase of analysis accuracy by decomposition of time period into sub periods and determination of time granulation. For the use of describing such dependency, the Virtual Time Factor was defined, which relates to relation of virtual and real time, according to the following formula:

$$VTF = \frac{V * I}{d}, \tag{1}$$

where d determines number of real time days, V determines number of virtual time days and I is a number of time intervals taken into consideration for monitoring of users' activeness. In example, if we assume that virtual day is equal to real day VTF=1. If during single real day, we can identify three virtual days, in example 3 logins per day can happen in the morning, afternoon and evening then VTF=3. Time lapse in virtual world has a different scope, and each virtual day for an individual user is initiated at the moment of logging into a system. Such approach enables monitoring the lapse of both types of time for each user and determining a level of activeness. For the purpose of data analysis consistently with assumed time granulation at the t moment for every *i-th* user the S_t snapshot is generated in parameters that are monitored in an individual time period. A set of snapshot parameters depends on parameters and abilities of a system and aims to present the dynamics of system usage and changeability of user's determined features. The m determination was assumed of $P^i_{s,t}$ parameters for every *i-th* user , that $i = 1,...,n$ for every type of s parameter that, $s= 1,...,m$ and for every t time moment the aggregated parameter values are determined in relation to previous periods, as well as dynamics of change $D^i_{1,t}$ in relation to previous period, according to formula:

$$S_t = \{\{P^i_{1,t} = \sum_{k=t-1}^{t_0} P^i_{1,k}\ D^i_{1,t} = P^i_{1,t} - \sum_{k=t-1}^{t_0} P^i_{1,k}\},\ldots,\{P^i_{n,t} = \sum_{k=t-1}^{t_0} P^i_{n,k}\ D^i_{n,t} = P^i_{n,t} - \sum_{k=t-1}^{t_0} P^i_{n,k}\}\} \qquad (2)$$

Analysis of individual parameters' changeability enables the determination of characteristics of audiences. Parameters can include social characteristics, user's activeness, a number of loggings and use of particular system functions. On the basis of collected snapshots from a user's account and a presence in a service, the relation of real time to virtual time is determined. Real time has a constant lapse with division into time intervals determining its granulation. For each user there can be determined characteristics, as well as dependency of virtual time on real time, and there is a possibility to model in such a way a dynamics of service usage. Time dependency between real and virtual time is a measure of user's engagement and dynamics of service usage. For the purpose of sequence analysis, the methods focused on measurement of similarity between sequence A and B can be introduced in a form of S similarity (A,B), which can be determined with a use of available methods. To evaluate similarity, time warping distance method was applied. For two series $X(x_1,x_2,\ldots,x_n)$ i $Y(y_1,y_2,\ldots,y_m)$ with lengths respectively n and m and M matrix is defined as interpoint relation between series X and Y where M_{ij} element indicates the distance $d(x_i,y_j)$ between x_i and y_j. Dependency among a series is determined by time warping path. The algorithm determines warping path with the lowest cost between two series in accordance with the formula [Rabiner, Juang 1993]:

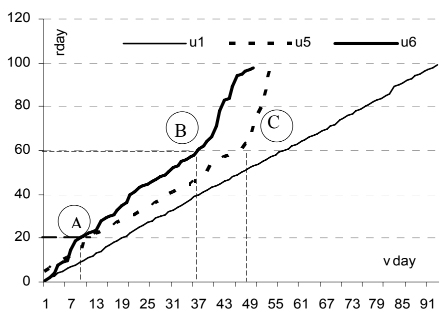
$$DTW(X,Y) = \min_W \{d_k, w = \langle w_1, w_2, \ldots, w_k \rangle\} \qquad (3)$$

where $d_k=d(x_i,y_j)$ indicates distance represented by $w_k=(i,j)$ on W path. A series with a higher level of similarity can be better compared because of alignment and dependencies resulting from dynamic time distance. To determine their relation to real time, the dynamic time wrapping was used. In the proposed procedure, the users' sequences can be compared to distinguish similar behaviour and compared to the pattern of ideal time lapse similar to real time. In the next part the research was conducted in real system, and a particular area of applications was identified.

EXPERIMENTAL RESULTS

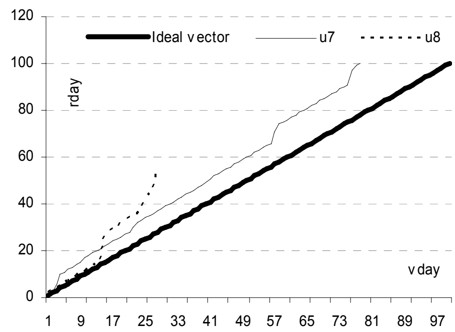
The proposed approach was verified on real data and was integrated with online system databases. In the scope of research, the analysis was made of the similarity between real and virtual time during the usage of social service realized in a form of virtual world. 16,165 users who at the time of research created accounts were taken into consideration. The more detailed analysis was made in relation to an input set of 4,410 unique users whose activeness was registered at least for 10 real days. Users were divided into 9 groups depending on the length of log- in sequences. The maximum number of virtual days is equal to the number of real days, and in the analysed case it amounted to 100. Granularity for time and virtual time factor was adjusted at 1:1, and real day is equal to virtual one. For particular users, changeable dependencies of two dimensions of time can be observed. Chart 1 shows the time dependency for real time R_d and virtual one V_d. On axes x and y, real and virtual days were marked. For u_1 user there is a consistency of measures, and high frequency of system usage is presented. User u_6 in the initial period increased a distance between real and virtual time, and his V_d in point A equals 12, and R_d equals 20. In the following days the distance is slightly changed to V_d=19, and there is a stability up to point B where a distance rapidly increases and eventually equals V_d=49 and R_d=100. For u_5 user, stability occurs from V_d=7 up to point C where distance rapidly increases between dimensions. Dependencies among users can be a basis for segmentation depending on behaviour similarities. The analysis was made including similarities and a distance to ideal vector, which represents a state of maximum engagement. In such a situation V_d is consistent with R_d. The exemplary dependency is shown in Chart 2. The distance and similarity to the ideal vector can be an engagement measure, and it enables detection of users with determined characteristics.

Chart 1. Relative users' behaviour characteristics u_1, u_5 and u_6.



Source: own calculations

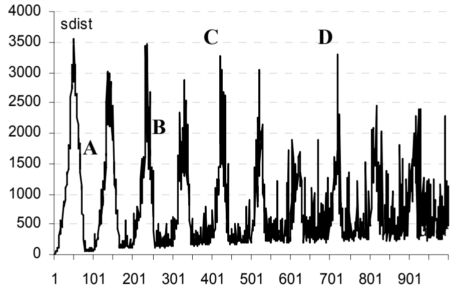
Chart 2. Users' characteristics u7 and u8 in relation to ideal vector



Source: own calculations

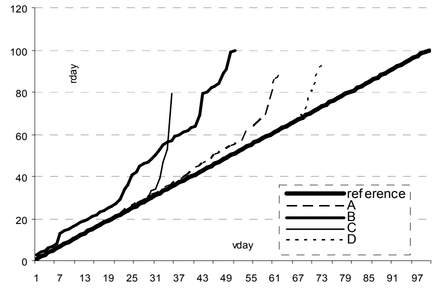
A similarity to ideal vector results not only from the sequence length, but also from the character of changes and lag of virtual time vector. Such characteristics are included by dynamic time wrap method that enables, among others, a presentation with comparison of sequence time-lag. Chart 3 illustrates in detail the distance for sequences from 1-1000 (window W_1 from Chart 4) and distances from ideal vector. Point A represents user u92 with dtw=68 and sequence length 74. Point B user u_{183} with distance 574 and sequence length 64/100 elements. Point C shows user u332 with DTW=2875 and sequence length 52. Point D denotes user u719 and dtw=3294 and sequence length= 36. Chart 4 is presented for selected points in relation to the ideal vector representing full consistency of measures V_d and R_d.

Chart 3. Detailed distance characteristics for users u_1-u_{1000}



Source: own calculations

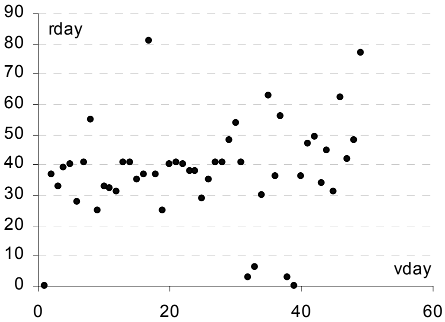
Chart 4. Distance from reference pattern



Source: own calculations

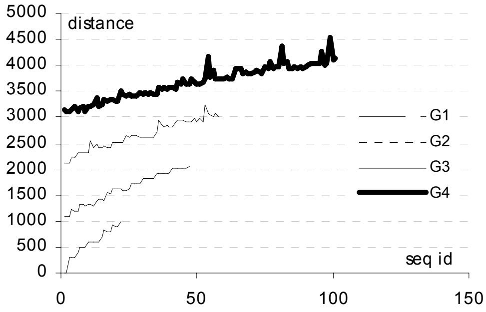
In the next step the users were divided into groups depending on the sequence length with incrementing every ten. In group g1 there were identified users with sequence length of 90-100 elements, in the second group from 80 to 90 up to group 9 in which the scope was from 10 to 20 elements. In particular groups the number of users was identified according to *card*(g_1)=22, *card*(g_2)=47, *card*(g_3)=63, *card*(g_4)=101, *card*(g_5)=133, *card*(g_6)=241, *card*(g_7)=398, *card*(g_8)=852, *card*(g_9)=2553. Chart 5 illustrates the similarity alignment between the longest sequence in individual set and other series in set g_2. Chart 6 illustrates the distance of individual groups to ideal vector.

Chart 5. Similarity alignment of series for set g_2



Source: own calculations

Chart 6. Distance alignment for elements from groups g1-g4



Source: own calculations

Analyses of dependencies of virtual and real time can be used for early detection of active users, and for estimation of the possibility of conversion on permanent users who often participate in the service. The presented idea and conducted research in relation to sequence of visits in the scope of real and virtual time show the possibility of behaviour pattern analysis with a use of this approach and detection of users with similar characteristics. Introduction of quantitative similarity measures and determination of factors influencing the occurring tendencies can broaden the existing data analysis methods and determination of audiences' parameters. It shows new approach to web data analysis based on dynamic time wrap algorithm and virtual/real time dimensions. In the next phase the research can be extended by time granularity analysis and influence of time patterns on the technical and economic parameters describing users' behaviour in the service. The scope of application of the presented concept is quite broad and enables the implementation of the monitor systems focused on behaviour examination and analysis of current tendencies and changes in audiences' preferences. The knowledge of the behaviour pattern enables the possibility of providing the recipients in a given group with marketing actions and integration of loyalty programs to increase the engagement of users classified in a group with a high risk of withdrawal from service usage, as well as to provide adequate offers to users with the highest levels of engagement.

SUMMARY

Together with the increase of Internet systems' complexness and increasing competition in this sector, there is a bigger demand for introduction of new data analysis methods, and examining behavior of users participating in Internet services. The proper recognition of needs and tendencies provides the basis for making rational decisions and better adaptation of online services to users' needs. High dynamics and changeability of users' behavior show the need for implementation of solutions that takes time characteristics into account. The solutions presented so far in the literature did not include two-dimensional approach towards data character and time series. The presented approach enables quantitative estimation of users' activity with the use of reference sequences and users' segmentation on the basis of time conditioning. The applied methods of dynamic time wrapping enable quantitative presentation of similarity including usage patterns. The presented research proves the possibility of using this approach in different fields and initiating new research areas focused on recognition of audiences' characteristics with bi-temporal representation.

REFERENCES

Acar, A.S., Polonsky, M. (2008) Online Social Networks and Insights into Marketing Communications, Journal of Internet Commerce, Volume 6, Issue 4, pp. 55-72.

Antunes, C.M., Oliveira A.L (2001) Temporal data mining: an overview. In: KDD Workshop on Temporal Data Mining, pp. 1-13, San Francisco.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O. (2004) Hourly Analysis of a Very Large Topically Categorized Web Query Log. In SIGIR 04, July 25- 29, Sheffield, pp. 321-328, South Yorkshire, UK.

Chau, M., Zeng, D., Chen, M., Huang, M., Hendriawan, D. (2003) Design and evaluation of a multi-agent collaborative Web mining system. Decisision Support Systems, vol. 35, no. 1, Elsevier Science Publishers, Amsterdam, The Netherlands.

Chun-Ling, Z., Zun-Feng, L., Jing-Rui, Y. (2010) The Application Research on Web Log Mining in E-Marketing. In: e-Business and Information System Security (EBISS), 2nd International Conference, 22-23 May, pp. 1- 4, Wuhan, China.

Dick, A.S., Basu, K. (1994) Customer loyalty: toward an integrated conceptual framework. Journal of the Academy of Marketing Science Spring , Volume 22, Issue 2, pp. 99-113.

Gomes, M.F., Canuto, A.M. (2006) Carcara: A Multi-agent System for Web Mining Using Adjustable User Profile and Dynamic Grouping. IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 187-190.

Gürsel, A., Sen, S. (2009) Producing Timely Recommendations From Social Networks. Proceedings of 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Decker, Sichman, Sierra and Castelfranchi (eds.), May, 10-15, Budapest, Hungary.

He, J., Chu, W.W. (2010) A Social Network Based Recommender System, Annuals of Information Systems (AoIS). Special Issue on Data Mining for Social Network Data, Volume 12.

Hernández-Ortega, B., Jiménez-Martínez, J., Martín-DeHoyos, M.J. (2008) Differences between potential, new and experienced e-customers: Analysis of e-purchasing behaviour. Internet Research, Volume 18, Issue 3, pp. 248-265.

Herring, S.C., Scheidt, L.A., Wright, E., Bonus, S. (2005) Weblogs as a bridging genre, Information Technology & People. Volume 18, Issue 2, pp.142-171.

Hu, X., Yin, Y., Zhang, B. (2007) Mining Temporal Web Interesting Patterns, International Conference on Computational Intelligence and Security (CIS). pp. 227-231.

Jung, J.J. (2005) Visualizing Recommendation Flow on Social Network. Journal of Universal Computer Science, Volume 11, Issue 11, pp. 1780-1791.

Karahanna, E., Straub D.W., Chervany, N.L. (1999) Information technology adoption across time: a cross-sectional comparison of pre-adoption and post-adoption beliefs. MIS Q., Volume 23, Issue 2, pp. 183-213.

Kazienko, P. (2009) Mining Indirect Association Rules for Web Recommendation. International Journal of Applied Mathematics and Computer Science, Volume 19, Issue 1, pp. 165-186.

Lu, H.P., Wang S. (2008) The role of Internet addiction in online game loyalty: an exploratory study. Internet Research, Volume 18, Issue 5, pp. 499-519.

Marciniak, R., Budnarowska, C. (2009) Marketing Approaches to Pop Up Stores: An Exploration of Social Networking. In: 15th Conference of the European Association of Education and Research in Commercial Distribution (EAERCD), 15-17 July 2009, University of Surrey, England.

Marsili, M., Slanina, F., Vega-Redondo, F. (2004) Dynamics of social networks (the rise and fall of a networked society). Proceedings of the National Academy of Sciences, Volume 101, pp. 1439-1442.

Newman, M.E.J. (2006) Finding community structure in networks using the eigenvectors of matrices. Physical Review, Volume 74, Issue 3, pp. 798-817.

Newman M.E.J., Girvan M. (2004) Finding and evaluating community structure in networks. Physical Review, Volume 69, Issue 2, pp. 1123-1138.

Rabiner, L. R., Juang, B. (1993) Fundamentals of speech recognition. Prentice-Hall.

Roddick, J.F., Myra S. (2002) A Survey of Temporal Knowledge Discovery Paradigms and Methods. IEEE Transactions on Knowledge and Data Engineering, Volume 14, Issue 4, pp. 750-767.

Tsai, H., Huang, H. (2007) Determinants of e-repurchase intentions: An integrative model of quadruple retention drivers, Information & Management. Volume 44, pp. 231-239.

Venkatesh, V., Morris, M. G., Davis, G. B., Davis, F. D. (2003) User acceptance of information technology: Toward a unified view. MIS Quarterly, Volume 27, Issue 3, pp. 425-478.

Wagner, C., Bolloju, N. (2005) Supporting knowledge management in organizations with conversational technologies. Journal of Database Management, Volume 16, Issue 2, April, pp. 1-8.

Wangenheim, F., Bayon, T. (2004) Satisfaction, loyalty and word of mouth within the customer base of utility provider. Journal of Consumer Behaviour, Volume 3, Issue 3, March , pp. 211–220.

Zhang, Y., Jansen, B.J., Spink, A. (2009) Time series analysis of a Web search engine transaction log. Information Processing and Management, Volume 45, Issue 2, pp. 230-245.