Web User Behaviour Prediction

Ondrej KAŠŠÁK*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
ondrej.kassak@stuba.sk

Web sites represent a repositories of huge amount of human knowledge. Early sites contained static content, which was represented mainly by texts. As the Web is becoming more dynamic nowadays [2], the text content was enrichted by other data formats, mostly by the multimedia ones as images, videos, music, but also the interactive elements changing with time and based on user actions. These ones are very difficult to represent when modelling users' behaviour.

With the Web 2.0, web sites got also the semantic concept with metadata used to describe elements' content for needs of various machines, which use these information to increase the usefulness of interaction with the site for the user.

Also many of web sites stop to offer the same content uniformly for every user and they become personalized. This means that content user actually gets from web site is adjusted individually for his/her preferences or actual needs. To be able to improve and potentially personalize the system content, we have to know as much information about users' characteristics and browsing habits as possible and model user's behaviour.

Information about web browsing can be extracted from different sources [1], with various complexity and in dependency to actual source they often need to be pre-processed. There exist three basic kinds of data describing user's behaviour [2]:

- *Web-structure* – web site can be represented as an oriented digraph consisting of vertexes (pages with unique URLs), oriented edges (hyperlinks between pages) and vertexes' characteristics (actual content of pages). This notion however covers the web pages with static content and unique URLs, but it cannot fit to dynamical sites which are continuously updated. For dynamic content, there is needed to find out new ways of representing the web sites structure.
- *Page content* – content of pages which the users interact with. Page semantic content is typically extracted by Natural Language Processing approaches and is used to mainly to estimation of similarities between pages. Similarities are in

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2015 PeWe Workshop, April 11, 2015, pp. 47–48.

static web sites based on text similarity measures, while in more dynamic sites they take into account the also time, because of content evolving in time.

- *User session* – user visit (click stream) on a web site can be represented as a browsing trajectory called the session. It can be logged directly or additionally reconstructed from a logs.

Based on all web site users' sessions, we are able to identify general web browsing patterns and the session characteristic as the typical session range, duration or the average amount of pages seen. We can use them when modelling user's behaviour and want to know the characteristics in which he/she differs from average. Based on them we are then able to describe the user and his/her expertise level, estimate probable future behaviour or to find the similar users or interesting content to recommend to him/her.

Possibility to estimate user's future behaviour can be very helpful in various situations, because it gives us the advantage to react to imminent consequences in advance. As example we can mention the situation when user leave the web site very soon. If we are able to predict this state, we can offer him/her an interesting content to keep him/her in system longer. In case of commercial systems as for example e-shop, this increases the chance that the user will buy something.

Task of user behaviour prediction is based on identification of attributes that can have influence on behaviour prediction and consequently of learning the importance of them for the task. Identified attributes can be assigned into several classes based of their origin. As mentioned above, there are three basic kinds of data describing user's behaviour. Web structure metadata can be used for example in sites with pages somehow logically ordered or grouped into hierarchies (topics in news, chapters in e-learning, categories in e-commerce). For example when user finish reading some chapter in e-learning system, there is higher chance he will leave than if he read the middle of chapter. Information about page content can be also used as prediction attributes – most important words or the page topic tell us what user read on the site and we can say how many similar pages we are able to offer to him. The third source of information is the user session. The knowledge of how many pages user saw in the actual session, how much time he/she spent there or how much his/her actions vary from the average behaviour are very important and useful.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. 009STU-4/2014.

References

[1] Liu, C., White, R. W., Dumais, S.: Understanding web browsing behaviors through Weibull analysis of dwell time. *33rd int. ACM SIGIR conf. on Research and development in inf. retrieval (SIGIR '10)*, (2010), pp. 379-386, ACM, USA.

[2] Roman, P. E.: Web User Behavior Analysis. PhD Thesis. UNIVERSIDAD DE CHILE., (2011)

[3] Tao Y.H., Hong T.P., Lin W.Y., Chiu W.Y.: A practical extension of web usage mining with intentional browsing data toward usage. Expert Syst. Appl., 36(2):3937–3945, (2009).