An Adaptive Cross-Site User Modelling Platform for Cultural Heritage Websites

Maristella Agosti[1], Séamus Lawless[2], Stefano Marchesin[1], and Vincent Wade[2]

[1] Department of Information Engineering,
University of Padua, Italy
stefano.marchesin@dei.unipd.it, maristella.agosti@unipd.it
[2] Department of Computer Science and Statistics,
Trinity College Dublin, Ireland
seamus.lawless@scss.tcd.ie, vincent.wade@scss.tcd.ie

Abstract. This paper discusses an adaptive cross-site user modelling platform for cultural heritage websites. The objective is to present the overall design of this platform that allows for information exchange techniques, which can be subsequently used by websites to provide tailored personalisation to users that request it. The information exchange is obtained by implementing a third party user model provider that, through the use of an API, interfaces with custom-built module extensions of websites based on the Web-based Content Management System (WCMS) Drupal. The approach is non-intrusive, not hindering the browsing experience of the user, and has a limited impact on the core aspects of the websites that integrate it. The design of the API ensures user's privacy by not disclosing personal browsing information to non-authenticated users. The user can enable/disable the cross-site service at any time.

Keywords: cross-site information needs, cross-site user modelling, web information exchange, user model provider

1 Introduction and Motivations

The exponential increase in Web usage has allowed modern societies to access, create, manage and distribute massive amounts of information. This fact has led to an increase in the rate at which information is created and consequently uploaded to the Web. The phenomenon is known as 'information explosion' and results in an increased difficulty in organising digital information to meet users' information needs.

A variety of systems have been created to try to address the problem by assisting users' information needs. These systems include, among other tools: keyword-based search engines, recommender systems, and web personalisation techniques, which adapt different aspects of the web experience to the needs and preferences of the individual user. Techniques like personalised information retrieval, where the search result list is re-ranked based on the user's individual search history, or social graphs[3] have been widely adopted.

[3] Social graphs '*socially connect*' users with content and products that 'peers' like. [3].

Within this context, however, most current approaches are unable to assist users in more complex conditions than just providing results for simple information needs, such as providing recommendations for a retail website. For example, a user's information needs that span different subject domains from different independently hosted websites represent a challenge which these "traditional" techniques cannot answer. In addition, the cross-site browsing process carries with it two phenomena that are known as 'lost in hyperspace' and 'information overload', which can both negatively affect the fulfilment of a user's information needs [1,2]. Therefore, to support the user and prevent them from being affected by these phenomena, a browsing experience is required that in a unified manner exploits all the content of the different websites the user is browsing/has browsed to tailor the content to be provided. In this way, although the traditional method of browsing is left unchanged, i.e. when a user browses individual websites across the Web moving from one to another, the conceptual vision of the browsing experience can be redefined as a unified and seamless browsing experience, not simply within, but also across different websites. Even websites, and consequently web publishers, would benefit from this new way of seeing the browsing experience, by gaining more insight on each individual user that landed on them and therefore being able to provide better content to users, thus prolonging their stay in the website.

Clearly there is a need for a consistent cross-site support mechanism that ensures effective assistance to users in the websites they browse across. A concrete representation of such a cross-site support mechanism, which is presented in this paper, is a third party cross-site user modelling service. This service is based on a user model provider, held and maintained by a single website, that is able to take specific aspects of each user together from different websites in order to provide cross-site information to target websites which can subsequently use it for personalisation. In addition, it has to ensure that the user is able to freely browse without any limitation or control, by implementing non-intrusive (implicit) tracking methods and allowing the user to decide when to activate the information exchange mechanism. Moreover, it is also necessary to investigate limited-impact techniques allowing information exchange mechanisms to be introduced to websites with a limited effort. Furthermore, the service should limit the flow of user's information from the third party user model provider to target websites, thus honouring users' privacy needs.

The paper is organized as follows. Section 2 reports on related work. Section 3 presents the proposed cross-site user modelling approach. Section 4 presents the service architecture. Finally, Section 5 reports on conclusions and future work.

2 Related Work

Web personalisation can be defined as any action that tailors the web experience to a particular user, or set of users, by providing the information users want or need without expecting them to ask for it explicitly, as described in [7].

Therefore, it is clear that a key challenge lies in providing web personalisation techniques that assist users across the Web and not only on isolated websites. To address this problem, web personalisation methods have to balance the needs of both website users and web publishers. For website users, the web personalisation techniques have to provide assistance across different websites, but in doing so honour the user's privacy needs and browsing freedom [6]. For the web publisher, the web personalisation techniques have to ensure simple and cost effective integration of web personalisation to existing websites and honour the website owner's control over the website as argued in [9], which has been a work of interest from which this project differs in some fundamental aspects highlighted in Section 4. Figure 1 sketches the concept of Cross-Site Personalisation (CSP). Based on this, it can be argued that a cross-site approach requires influencing both areas simultaneously and in real-time by creating a *state of equilibrium* in which the needs of the user and the web publisher are balanced.

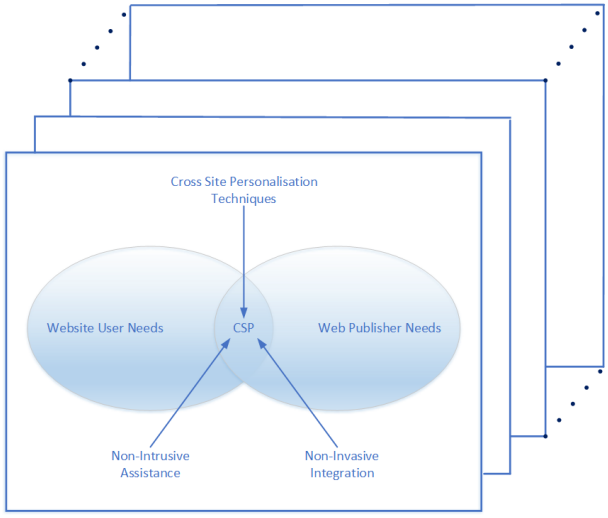

Fig. 1. Cross-Site Personalisation Concept.

In order to integrate a cross-site approach to websites, limiting the integration impact to websites, information exchange techniques at run-time can be investigated. Such techniques have their main challenges in understanding the interdependencies within a website. To address this integration challenge Web Information Systems (WIS) can be used, which allow the implementation of an entire website as out-of-the-box deployment [5]. However, WIS are traditionally proprietary software products that limit extensibility without the vendors help and/or approval. Therefore, to overcome this limitation Web-based Content Management Systems (WCMS) can be used instead, which allow a simple and more flexible implementation based on their open source nature. The main

feature of WCMS is the extendible framework, which allows external modules to influence different levels of the functional layer of the website without the need for re-design or re-deployment. This pluggable nature of WCMS allows for low impact personalisation approaches in the form of external modules that can be enabled/disabled by web publishers at ease. Furthermore, applying the cross-site service to WCMS provides the additional benefit of giving web publishers control in deciding which area(s) of the website should be affected by the cross-site user modelling platform and also in deciding how much the extension should influence the overall website look and feel. For example, a web publisher may decide to make the service visible in the homepage or only in specific sections of the website, depending on their needs.

Regarding the research area of 'user modelling', which is a wide and complex area, the following discussion is mainly focused on implicit user models in order to comply with the non-intrusiveness notion adopted for the cross-site service presented in this paper [4]. Within this area, user models that allow high-level abstraction are often referred to as user profiles. User profiles can be defined as a subclass of user modelling, less sophisticated and more suited for applications that require a more general abstraction of information needs [8].

To allow the user to gain a deep understanding in the meaning of the extracted terms and to overcome the problem of polysemy semantic, term networks can be used. Semantic networks usually consist of a term structure, which entails an order or relationship.

3 Cross-Site User Modelling Approach

The domain chosen for the use-case presented in this paper is cultural heritage. Within this domain the Virtual Research Environment (VRE) for the Digital Humanities of the CULTURA[4] project has been investigated. This VRE supports users with different levels of experience to use a variety of tools to interact with a number of cultural heritage collections. The study of this work inspired us to envisage a parallel and complementary approach. This approach provides relevant information to users that attempt to answer cross-site information needs within its browsing space. In those situations where the user's need spans topics that are not confined to a single website, the introduction of the proposed cross-site service might improve the effectiveness of the website personalisation, along with the user's level of satisfaction. An example of this could be an overarching user's interest in the living conditions of the Irish middle class during Irish rebellions, which cannot be addressed by a single website of the CULTURA VRE but requires the user to navigate across different websites of the VRE browsing space. Hence, with the introduction of the cross-site service the user model could gain a higher precision in those topics that are more relevant to the user and, therefore, the website the user is browsing might be able to provide more tailored personalisation to help address the user's cross-site information needs.

[4] `http://www.cultura-strep.eu/`

The use-case process identified is as follows: (1) The user lands on a website related to an information gathering task in the cultural heritage domain; (2) The user authenticates a first time with the third party service; (3) The website tracks all user activities in the webpages along with the relevant text entities identified by a term identification component; (4) The user triggers the information exchange function, in anticipation of subsequent personalisation by the target website, which provides relevant user data (based on the selected communication pattern) to the website and newly tracked user information to the service; (5) The user surfs to a second website and, depending on whether they are already authenticated or not, authenticates or directly triggers the information exchange function, which should provide more tailored information to the website; (6) Steps (1)-(5) can then be re-iterated many more times, without a strict order of execution.

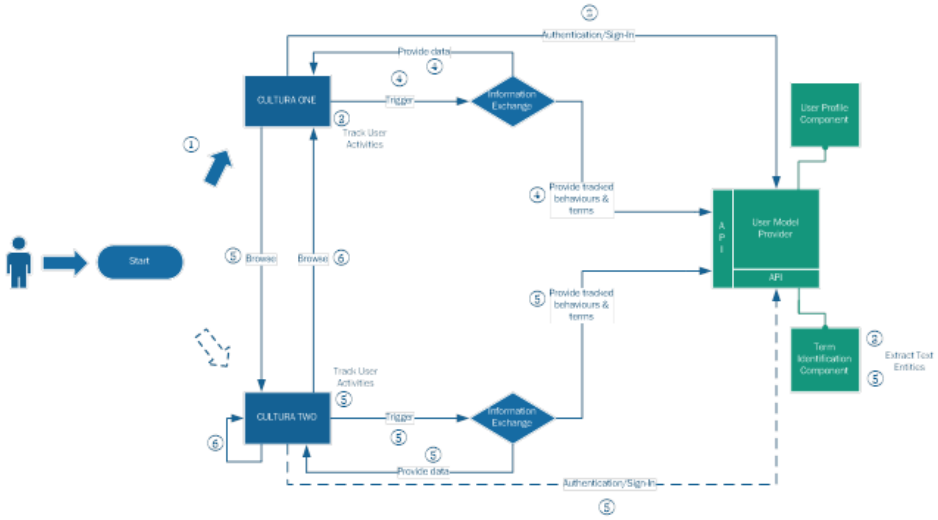

Fig. 2. High-Level Possible Use-Case Process.

The use-case approach described above requests the following features in order to be applied: (1) websites enable/allow third party sign-in/authentication by using OAuth[5]; (2) the website needs to communicate with the user model provider. The communication includes: (a) sending of content browsed by users for term identification to a third party service, (b) enable/allow browsing behaviour tracking, (c) sending tracked user activities and extracted text entities to the user model provider; (3) the website has to use custom-built WCMS module extensions in order to access the cross-site information exchange service RESTful API; (4) user authentication with the third party service is required in

[5] `http://tools.ietf.org/html/rfc5849`

order to enable the information exchange mechanism and to protect users from unregulated treatment of their data.

The following section introduces the architecture of the proposed cross-site user modelling platform.

4 Cross-Site User Modelling Architecture

According to the components depicted in Figure 3, the high-level service architecture is presented below, highlighting, where necessary, the differences from the work in [9].

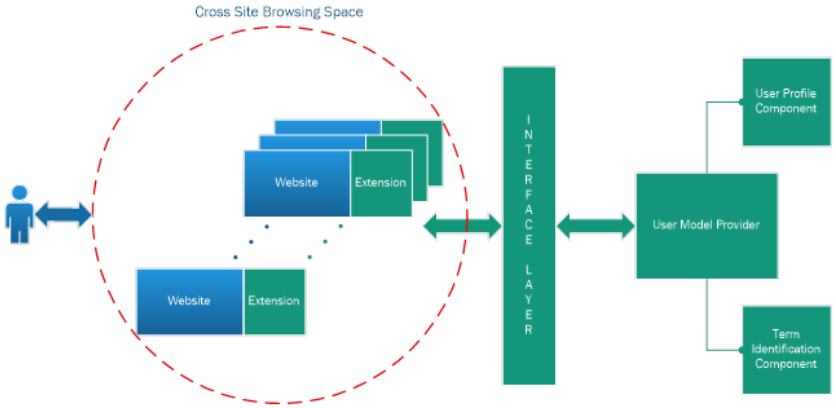

Fig. 3. High-Level Service Architecture.

Term Identification Component: The main purpose of the term identification component is to identify text entities related to the body of the current webpage the user is viewing. Text entities indicate the meaning of the underlying content from websites belonging to the cross-site browsing space. A text entity can be considered any set of useful information connected to an extracted content-related term. Examples of this term-connected information can be the topic, the confidence level (which is the probability that the extracted term refers to the connected topic) and the resource reference. Text entities are retrieved through an external term extraction tool. An additional responsibility of the term identification component is to ensure the creation of a shared conceptualisation of the user's cross-site browsing space. The shared conceptualisation is represented as a text entity space and it is based on the contents the user has browsed within the websites of the cross-site browsing space. In this sense, taxonomies and ontologies used by external term identification tools play a fundamental role in defining the characterisation of the shared conceptualisation. Generic ontologies, like DBpedia[6], will lead to a shallower shared conceptuali-

[6] `http://wiki.dbpedia.org/`

sation than specific domain ontologies. Therefore, depending on the domain of application, different ontologies have a different impact over the granularity and quality of the conceptualisation. However, in order to extract terms and create such a shared conceptualisation, the term identification component needs access to the openly accessible contents on the different websites. Within the interface layer of this architecture, it is proposed that the website sends at run-time the contents the user is currently browsing to the term identification component. Hence, the external tool returns to the website the text entities extracted.

User Profile Component: The user profile consists of text entities related to the current task of the user and that were identified by the term identification component. In addition, activities the user has conducted on the content are also stored, helping to understand which of the contents browsed by the user, and therefore which of the topics inferred by text entities, are more relevant. These activities can include mouse clicks, scrolls, cut & paste operations etc.

Interface Layer: The interface layer provides an abstraction from the specific implementation of the different websites within the user's cross-site browsing space. It does so by implementing a RESTful API (Representational State Transfer[7]) that facilitates the communication between the user model provider and the websites the user is browsing. A key responsibility of the API is to ensure that the interface with the user model provider is browser independent and does not depend on a specific technology stack. Furthermore, it should ensure fast and accurate interconnectivity between the interfacing websites and the cross-site information exchange platform.

Web-based Content Management System Module Extensions: WCMS module extensions allow a simple and limited-impact integration of cross-site information exchange techniques into existing website implementations. The responsibility of the WCMS module extensions is twofold: (1) To facilitate communication between the website and the API of the cross-site user modelling platform and (2) to provide non-intrusive information exchange techniques to the user, within the website the user is currently browsing. The former (1) is achieved thanks to the interface layer and its ability to abstract, through the HTTP methods, from the specific implementations of the different websites. The latter (2) is achieved through the use of implicit tracking methods for the user's activities within the website and the implementation of a website tool (i.e. a button) that allows the user to decide whether to activate the information exchange mechanism or not. Depending on this, the text entities returned by the term identification component are kept within the website until either the user activates the information exchange service or their session ends. In this way, as opposed to [9], the website does not have to send all its contents to external services a priori, but rather to send those browsed by users only. Furthermore, by sending text entities to the cross-site service only when users activate the information exchange mechanism, the approach can be considered the basis for '*personalisation on demand*'. This allows websites (i.e. web publishers) to better

[7] `http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm`

preserve their contents and therefore to more easily accept the introduction of a third party cross-site user modelling service.

Cross-Site Browsing Space: The application of information exchange techniques to independently hosted websites introduces a cross-site browsing space, an example of which has been presented in Section 3. Within the browsing space target websites receive user data through information exchange techniques. The information exchange techniques are enabled through the aforementioned WCMS module extensions. Based on the back-end integration of the user model provider with the website, the user can enable/disable the information exchange service whenever they want. This allows the user to control the mechanism and ensures the user's privacy is honoured. The enabling/disabling of the service passes through the service's authentication system, in fact the user has to authorise the website to interface with the service via third party sign-in authentication with it. Hence, the website sends the cross-site service user's credentials and authorises the third party service to receive and send relevant cross-site information. In this way, the only authentication required is the one with the cross-site user modelling service, thus avoiding multi-log problems and not burdening the browsing experience of the user too severely.

Regarding information exchange techniques, several different methods have been designed. The techniques range from highly generic to highly specific, providing different levels of information enrichment for both the websites and the service. The former tend to be more satisfactory from a web publisher perspective, allowing a massive flow of user's information to leak from the service, thus ignoring or not sufficiently considering the user's privacy concerns. The latter, on the other hand, are more focused on preserving the user's privacy, thus avoiding the provision of huge amounts of information to websites or, in extreme cases, not providing it at all, tend to be more satisfactory both for the user, who sees their privacy more respected, and the service itself, as it doesn't give away the only real value it holds – user data.

For each technique, pros and cons related to privacy concerns were pointed out along with a technical review focusing on the effectiveness of the exchange. However, no claim was made regarding an ideal or optimal solution: all the techniques present both advantages and shortcomings that might be relevant depending on the context and the scope of the application that deploys them. Table 1 summarises the comparison of the exchange techniques in relation to their effectiveness. Hence, for each technique three aspects related to effectiveness are considered: volume, quality and granularity. Volume refers to the amount of data returned to the target website by each information exchange technique. Quality refers to the ability of each information exchange technique to provide tailored information to target websites, therefore exchange techniques that return huge amounts of information tend to have a lower quality in the information they provide. Granularity refers to the level of detail of the information returned to the target website, therefore information exchange techniques that return only entities or informed decisions tend to have a higher level of granularity. Each

aspect can take on the values "High", "Medium" and "Low", keeping in mind that a "Low" granularity refers to a high level of detail.

Information Exchange Techniques	Volume	Quality	Granularity
Privacy Insensitive Technique	High	Low	Low
Threshold Technique	Medium	Low	Low
Top-Feature Selection Technique	High/Medium/Low	Low/Medium	Low
Ranking Technique	Medium	Medium	Low
Entity-Oriented Technique	Medium	Medium	Medium
Activity-Oriented Technique	Low	Medium	Low
Knapsack Technique	Low	Medium	Low
Semantic Technique	Low	High	Low/Medium
Suggestion-Oriented Technique	Low	High	High

Table 1. Effectiveness Comparison.

5 Conclusion and Future Work

This paper presented a cross-site user modelling platform for information exchange techniques. The approach introduced serves as a starting point to address a gap between current web personalisation services and what should be a seamless browsing experience for users across independently hosted websites. Thus, the approach fits in the CSP context, through the use of a third party user model provider and WCMS extensions. Therefore, the main contribution of the research described in this paper is the enhancement of user profiles through the usage of a shared conceptualisation, which results from the aggregation of the text entities extracted from the websites belonging to the cross-site browsing space. In addition, the integration of limited-impact information exchange techniques at run-time contributed to the area of web engineering.

The architecture presented in this paper serves for the implementation of a service prototype which is still in progress and will be the subject of future work. The evaluation of this preliminary prototype focused on the effectiveness of the information exchange techniques designed. The first results extracted are encouraging and motivate the following two areas to be addressed in future work.

First, the natural direction this research work should take, in order to evolve, is the implementation of personalisation techniques. Hence, due to the semantic nature of the information stored in the third party user model, i.e. the text entities extracted by the term identification component, personalisation techniques that are able to understand and handle semantic data, such as semantic recommender systems, could be used. Second, user profile management and user scrutiny could be considered in order to enhance the user's trust and control over their needs. Therefore, to allow users to actively engage with their user profile, the cross-site service has to: (1) allow users to view terms relating to their cross-site information needs (model scrutiny); (2) enable users to add and delete

terms within the user profile; (3) provide insight on where the information was collected and used. The CULTURA project provided an initial form of model scrutiny, allowing the user to visualise their interests as a word-cloud where the terms could be enlarged or diminished by the user depending on the relevancy that the user gave them [10, 11] and this can be considered as a starting point for future work.

Acknowledgements

Stefano Marchesin would like thank Trinity College Dublin for the use of facilities during his stay in 2016. We thank the reviewers for their insights and comments on the earlier version of the manuscript.

References

1. Ahn, J.W., Brusilovsky, P., He, D., Grady, J., Li, Q.: Personalized web exploration with task models. In: Proceedings of the 17th International Conference on World Wide Web. pp. 1–10. ACM (2008)
2. Berghel, H.: Cyberspace 2000: Dealing with information overload. Commun. ACM pp. 19–24
3. Dieberger, A., Dourish, P., Höök, K., Resnick, P., Wexelblat, A.: Social navigation: Techniques for building more usable systems. Interactions
4. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: The adaptive web. chap. User Profiles for Personalized Information Access, pp. 54–89. Springer-Verlag (2007)
5. Isakowitz, T., Bieber, M., Vitali, F.: Web information systems. Commun. ACM pp. 78–80
6. Kay, J.: Scrutable Adaptation: Because We Can and Must, pp. 11–19. Springer Berlin Heidelberg (2006)
7. Keenoy, K., Levene, M.: Personalisation of Web Search, pp. 201–228. Springer Berlin Heidelberg (2005)
8. Koch, N., Wirsing, M.: Software engineering for adaptive hypermedia applications. In: 8th International Conference on User Modeling, Sonthofen, Germany (2001)
9. Koidl, K., Conlan, O., Wade, V.: Cross-site personalization: Assisting users in addressing information needs that span independently hosted websites. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media. pp. 66–76. ACM (2014)
10. Sweetnam, M., Siochru, M., Agosti, M., Manfioletti, M., Orio, N., Ponchia, C.: Stereotype or spectrum: Designing for a user continuum. In: the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH (2013)
11. Sweetnam, M.S., Agosti, M., Orio, N., Ponchia, C., Steiner, C.M., Hillemann, E.C., Ó Siochrú, M., Lawless, S.: User Needs for Enhanced Engagement with Cultural Heritage Collections, pp. 64–75. Springer Berlin Heidelberg (2012)