# COMBINING DESKTOP DATA AND WEB 3.0 TECHNOLOGIES TO PROFILE A USER

Vasileios Lapatas and Michalis Stefanidakis
*Department of Informatics, Ionian University, Plateia Eleftherias, Corfu, Greece*

Keywords: Web personalization, Personalized web sites and services, Web site classification, Learning user profiles.

Abstract: An interesting promise of Web 3.0 is the seamless integration of desktop and web spaces. Private data, locked until recently inside a user's computer, can lead to the intelligent generation of web content. This paper presents an idea of how desktop and web data can be integrated in creative ways in the World Wide Web. As a proof of concept, an application which can profile a user based on his bookmarks is being demonstrated. An existing web service is used to classify bookmarks, enabling thus platforms with limited processing power to perform the profiling process. Results gathered from the classification process indicate that even a generic untrainable and not fine-tuneable classifier can produce results with high accuracy. With accurate user profiles web content can be created in an intelligent way, enabling better Web 3.0 applications.

## 1 INTRODUCTION

Web mining is a key component of Web 3.0. With web mining, data on the web can be processed by computers for further use. Data used as input source to web mining can be contents of web pages (Web Content Mining), user activity (Web Usage Mining) or website structure data (Web Structure Mining).

All this information resides at server side and is constituted of the data the user, voluntarily or not, submits to web services' sites. On the other hand, an interesting aspect of Web 3.0, currently underdeveloped but with future potential, is the seamless integration of desktop and web data. This paper aims to show that desktop data residing on user's client side, can be used in the Web to create even more "intelligent" websites.

Recently, projects connected to the social web like MyTag (Franz, 2009) and Semantic Blogging (Möller, 2005) try to unlock the previously underutilized user's data. Other implementations suggest the usage of web technologies in a desktop environment (Aumueller, 2005) (Sauermann, 2005). Those technologies can make easier the web-desktop data integration. However, there appears no transparent use of desktop data in the web until now. Applications that use desktop data on the web always need user interaction and they usually have to store user files on the web.

The following sections show that with present technologies, desktop and web data integration can be achieved. As a proof of concept, an application was created that can profile a user based on the bookmarks saved in his browser. The application can serve as an API (Application Programming Interface) and the results of processing can be easily shared with other applications to enhance the user's experience, for example, to customize web data.

The rest of the paper is structured as follows: in the second section related projects are being presented and discussed. The third section describes the application and presents some results concerning it's accuracy. Privacy issues are addressed next and the final section lists future development ideas along with the authors' conclusions.

## 2 RELATED WORK

Over the years a few similar ideas via different approaches were presented, summarized in this section.

MyTag (Franz, 2009) presents user's data that are stored on the web as a cross web-based interface. MyTag exploits web services from various sites to access user data. The application presented in this paper is based on a similar idea but uses desktop data for user profiling and needs no user interaction.



Lapatas V. and Stefanidakis M.
COMBINING DESKTOP DATA AND WEB 3.0 TECHNOLOGIES TO PROFILE A USER.
DOI: 10.5220/0002806903500353
In *Proceedings of the 6th International Conference on Web Information Systems and Technology (WEBIST 2010)*, page
ISBN: 978-989-674-025-2

Semantic Blogging (Möller, 2005) uses desktop data for the needs of blogging activities. With Semantic Blogging a user can easily handle his own desktop data (contacts, calendars etc.). Semantic Blogging is the most closely related application to the presented one but again needs user interaction in order to be used.

WikSAR (Aumueller, 2005) is a web application that uses desktop data in a wiki environment (such as addresses, calendars etc). Although this project can use local data on the web, it doesn't process them by any means; it only presents them with a more elegant way of browsing.

Gnowsis Semantic Desktop (Sauermann, 2005) translates desktop data into semantic data for major operating systems. Although not web related, it supplies an easy way to access desktop data based on semantic analysis.

Automatic Bookmark Classification (Benz, 2006): This project has a lot in common with the one presented here, as it also classifies a user's bookmarks. However, there is no automated usage of the classification process; the user is prompted to accept the classification result or insert the results he believes that suit best.

Personalized Search (Teevan, 2005) studies the impact of web mining techniques on a search engine. They use of logs from previous searches as well as previously visited web sites to profile a user and return more accurate results. Only web data is being used, not taking advantage of desktop local data.

# 3 THE APPLICATION

As proof of concept for the aforementioned idea of exploiting desktop data on the web, an application has been developed using the Python programming language. This application is able to profile a user by using the bookmarks he has assigned to his web browser. This program has no user interface and the results are being acquired without any user interaction.

This application belongs to the Web 3.0 family, as it takes advantage of web services in order to detect what the user is really interested in. The web service being used is a classification service (URLclassifier, 2009). URLclassifier is a web service accepting a URL and returning the categories that are embedded in the relevant web page, as a result. After the application gathers the user's bookmarks, it uses this service to classify them into categories. Those categories are actually the user's fields of interest and can be used to profile the individual whose bookmarks were processed.

An on-line classification service is being used in order the local application not to consume a lot of processing power. Thus, the application can be used in mobile devices that do not have the processing power needed for the execution of a classification program. With the on-line service, full text of web pages is being processed. Another way of creating a light-weight classifier is to use only the URL of a web page as input to classification (Kan, 2005) (Baykan, 2009).

The application extracts the user's bookmarks and processes them with the help of the on-line classifier. Each of the popular browsers has a different way of storing it's bookmarks. Safari stores it's bookmarks in a .plist file while Firefox uses a .html file and Opera a plain text file. For everyone of the above browsers a different parser was developed, enabling the program's first phase to gather bookmarks from each one them. In a subsequent phase, the application sends the gathered bookmarks one by one as input to URLClassifier, in order to get the necessary results.

After the classification, all the results are stored into a list which can be consumed by a third party application or just be printed for debugging purposes.

## 3.1 Application's Accuracy

A number of computer science students was asked to submit their bookmarks to use them as input for the presented application, in order to test the application's accuracy with real life samples. Some samples contained a big number of bookmarks (greater than 100), while some others only few (less than 10). This was advantageous to the experimentation because it was possible to examine if the success percentage of the application varies between different input sizes.

After applying the classification process that was described in the previous section, the outcome was handed back to the users in order to evaluate the correctness of the classification results. From their answers a set of hit-miss counts was generated.

Figure 2 shows the five results of a single set with the maximum number of bookmarks. This particular set consists of 120 web sites and it was the largest set of those that were tested. The left bars indicate the successful categorization of the sites while the right bars indicate the false categorization of the sites. As someone can see in this set, the classification is quite accurate. To support this

speculation, Figure 3 shows the success rate of the top five categories with an average success rate of 89%. The total success rate of the whole sample (including the omitted results) is 72%.

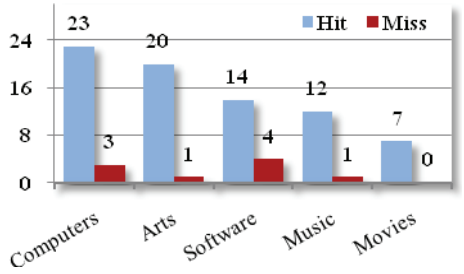


Figure 1: Top five results of the biggest set.

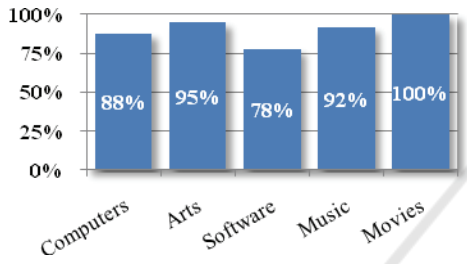


Figure 2: Success rate of the biggest set's top five results.

The application presented here works only for sites in English because URLclassifier service does not have support for other languages. Sites that are not in English are being ignored for now. In this particular set around 50 of the 120 sites didn't return any results and almost all of them were sites with non-English text (41% of the sample).

Let's take a look now to a complete different set, which consists of 22 bookmarks and was the smallest set available.

Once again, the results were quite accurate (Figures 3 and 4). The percentage of the success in the top five categories is 87,5% while the total success rate of the set is 87%. In this case those two values have a very small difference because almost all of the results are included in the top five categories (only 7 results are being omitted).

Like the first set, this one had also some sites that didn't return results because those sites were using the non-English language. (7 sites or 31% of the sample).

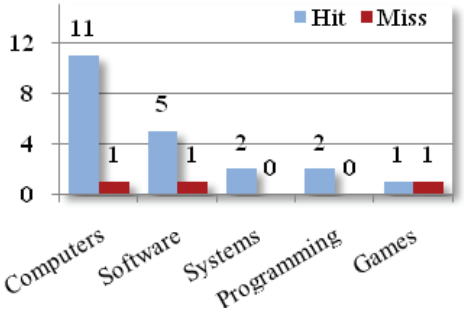


Figure 3: Top five results of the smallest set.

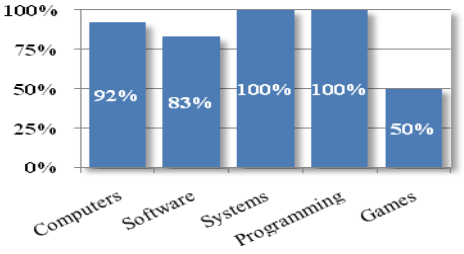


Figure 4: Success rate of the smallest set's top five results.

The figures shown before indicate that the classification success rate and the relevant user satisfaction are high. When the users were asked about the results of this application, most of the feedback was strongly positive and even the worst review was also positive.

## 4 PRIVACY ISSUES

In order to respect the user's privacy, whenever classification data is going to be sent to a Web 3.0 service, it is absolutely necessary not to allow the presented application to leak information about any of the user's bookmarks or personal files. The only information that should pass from the user to the server will be the classification results.

Even if classification only results are allowed to leave the user's computer, this action can be allowed only after user's agreement. A possible solution can be a license agreement. Unfortunately, the problem with license agreements is that very few people actually read them.

In order to increase user's awareness of the application actions and make sure that he understands completely what this application will do in his machine, another method is planned to be used. The first time the application is executed, an intuitive configuration window will be used to give the user the chance to review and control the data that will be processed, instead of just displaying

incomprehensible license terminology. The configuration dialog will appear only once and thereafter the application will be able to run transparently, without any further user interaction. To ensure the existence of a constant but at the same time unobtrusive warning to the user, small notification icons or messages, that do not require user response, may be used at the desktop.

Private data leakage prevention is a major issue in all desktop-web integration efforts and not specific to the application presented herein (Dunn, 1997) (Thuraisingham, 2002). It is obvious that a total solution provided by standard OS services is needed, as desktop and web spaces are getting fused within each other.

# 5 CONCLUSIONS AND FUTURE WORK

This paper investigates ways to utilize private data like user's bookmarks, locked until recently inside a user's computer for the intelligent generation of Web 3.0 content. An application, which generates the profile of a certain user based on the content of his browser bookmarks, was demonstrated. An external classifier service enables the execution of the application on platforms with limited processing capabilities like mobile devices and netbooks. The resulting figures indicate that with the usage of a generic non-trainable and non fine-tunable classifier it is possible to achieve satisfactory results in over 70% of cases in average and near 90% for top 5 categories in user's profile.

Future work will include the exploration of web site personalization mechanisms based on analyzed desktop data and the creation of suitable applications like programs that suggest to the user web sites that he might be interested in or bring people who have the same interests in touch automatically. On server-side, research will focus on search engines that can deliver more accurate results depending on the user profile.

# REFERENCES


Aumueller, D. and Auer, S., 2005. Towards a Semantic Wiki Experience - Desktop Integration and Interactivity in Wiksar. *In proceedings of the 1st Workshop on The Semantic Desktop, Next Generation Personal Information Management and Collaboration Infrastructure.*

Baykan, E., Henzinger, M., Marian, L. and Weber, I., 2009. Purely URL-based topic classification. *WWW '09: Proceedings of the 18th international conference on World wide web.*

Benz, D., Tso, K. and Thieme, L., 2006. Automatic bookmark classification - a collaborative approach. *In Proceedings of the 2nd Workshop in Innovations in Web Infrastructure (IWI2) at WWW2006.*

Dunn, M., Gwertzman, J., Layman A., Partovi, H., 1997. Privacy and Profiling on the Web. [online] [02 November 2009] <http://www.w3.org/TR/NOTEWeb-privacy>.

Franz, T., Dellschaft, K., Staab, S., 2009. Unlock Your Data: The Case of MyTag. *In FIS '08: Proceedings of 1st Future Internet Symposium.*

Kan, M. Y. and Thi, H. O. N., 2005. Fast webpage classification using URL features. *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management.*

Möller, K., Decker, S., 2005. Harvesting Desktop Data for Semantic Blogging. *In proceedings of Semantic Desktop Workshop at the ISWC.*

Sauermann, L., 2005. The Gnowsis Semantic Desktop for Information Integration. *In proceedings of IOA 2005 Workshop at the WM.*

Teevan, J., Dumails, S. and Horvitz, E., 2005. Personalizing search via automated analysis of interests and activities. *In SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.*

Thuraisingham, B., 2002. Data mining, national security, privacy and civil liberties. In *SIGKDD Explor. Newsl.*

URLclassifier, 2009. [online] [08 September 2009] <http:///www.urlclassifier.com>.