# AUTONOMOUS NEWS PERSONALISATION (ANP)

Mohammed Sharaf Al Zebdi and Tereska Karran
*Cavendish School of Computer Science, University of Westminster*
*115 New Cavendish Street, London, W1M 8JS, U.K.*

Keywords: Personalisation, Web Usage Mining, Data Mining.

Abstract: This research explores some of the directions for improving the performance of personalised web usage mining applications. The study uses ANP (Autonomous News Personalisation) to provide personalised news to online newsreaders according to their interests. This is achieved within an intelligent web browser which monitors users' behaviour while browsing. Web usage mining techniques are applied at the site's access log files. These are first pre-processed, and then data-mined using specific algorithms to extract the interests of each user. User profiles are created and maintained to store users' interests. User interests within the profile are ranked according to their reading frequency of news items ranked according to category and location. Profiles are refined continuously and adapt to users' behaviour. Besides being adaptive and completely autonomous, the system is expected to improve on existing performance in news retrieval and to provide higher level personalisation. A system prototype has been implemented and tested using SQL Server 2005 to pre-process logs, data-mine cleaned data, and maintain user profiles. The main system tasks can be demonstrated with further work to address all the issues.

## 1 INTRODUCTION

The amount of text data on the web is growing enormously, while users' ability to keep up with it seems to be limited. Consequently, the concept of personalisation is growing in importance. One major growth data area is the news websites (Paliouras et al., 2006; Ardissono et al., 2000).

There are a large number of news websites, usually covering the top world news and aiming for general coverage of the majority interests. Because of the huge number of news items, and the diversity in people's interests, most people tend to read only specific types of news in specific locations. So, when log on, they waste part of their time scanning existing news and trying to locate items that match their interests. Some sites solve the problem by allowing users to specify their interests manually and log in to the system using 'credentials'[1]. Although this could potentially save users' time and effort, the user is still required to specify some data manually. Moreover, there are problems as users interests evolve continuously depending on world events.

ANP aims to provide users with news related to their specific interests directly without the need to neither specify them manually nor even log in using credentials. One way of achieving this is to develop intelligent news websites/browsers, which automatically identify users and monitor their behaviour as they are browsing. The aim of this approach is to profile the needs of each user and provide relevant news automatically. It is predicted that this process will have two benefits. Firstly, it will minimise users' time and effort on extracting news information. Secondly, it will make the process of browsing news sites more interesting and efficient. Furthermore, the process has applicability across different personalisation mechanisms.

ANP is designed to be applied on existing news websites. It consists of a data warehouse that collects data from web server log files through an ETL (Extract Transform Load) layer. The collected data is then pre-processed and data-mined to analyse users' behaviour and produce a profile for each type of user (user clusters will evolve over time). These profiles are maintained in a separate ANP database

[1] A credential is an electronic representation of user's identity (user name and password).



Sharaf Al Zebdi M. and Karran T. (2008).
AUTONOMOUS NEWS PERSONALISATION (ANP).
In *Proceedings of the Third International Conference on Software and Data Technologies - ISDM/ABF*, pages 263-267
DOI: 10.5220/0001875202630267

Table 1: Personalisation features achieved by different systems/studies.

| Feature / System | Automation | Adaptation | Performance | Matching User Requirements |
|---|---|---|---|---|
| **Google News (Google, 2007)** | partially autonomous | adaptive | high | Medium |
| **Yahoo News (Yahoo, 2007)** | non autonomous | non adaptive | high | Medium |
| **Adaptive User Profile for Filtering News (Singh et al. 2006)** | partially autonomous | adaptive | low | High |
| **Mining Web Logs of an On-line Paper (Batista and Silva, 2002)** | autonomous | non adaptive | high | Low |
| **ANP** | autonomous | adaptive | high | High |

which is used by web servers to filter news items with the aim of delivering personalised news presentations.

ANP aims to make online news systems fully autonomous including user identification, data collection and preparation, data analysis, building user profiles and news filtration. As well as providing adaptive high-level personalisation, the ANP should enhance performance on the client and on the server as well.

## 2 RELATED WORK

Great deals of academic research as well as commercial applications have been done in news personalisation. Most of them were trying to achieve four main features:

- Autonomy: how autonomous is the developed system? Is it completely autonomous or it requires some tasks to be done manually?
- Adaptation: is the system dynamic? i.e. does it change according to user's behaviour?
- Performance: how fast is the tasks of the personalisation process are accomplished?
- Matching user requirements: to what extent the delivered news matches user requirements?

Table 1 below compares between five different studies/systems, including ANP, in terms of achieving the aforementioned features. Those results are verified experimentally in section 4.

Overall, all of these studies/applications for managing news personalisation manage specific features well. However, none meets all the features required for personalisation in an efficient way. Some of them achieved high performance, but resulted with poor degree of personalisation and adaptation. Others produced high-level personalisation but suffered from low performance and high complexity. Therefore there is an urgent need to develop a system that compromises between those features, and results with an adaptive, autonomous and high-performance personalisation that matches user requirements in an acceptable degree.

## 3 PROPOSED SOLUTION (ANP)

ANP provides a personalisation service to news websites aiming to give users different news presentations that match their interests, while minimising processing. It uses web usage mining technology to monitor users as they are browsing news websites by modelling their needs based on analysis. The project aims to compromise between the features of personalisation discussed earlier. The approach applied involves a set of methods and techniques as follows:

- The first stage collects usage data collection from web access log files. This log data is pre-processed to remove noise (non-related data) and transformed where appropriate to be ready for analysis.

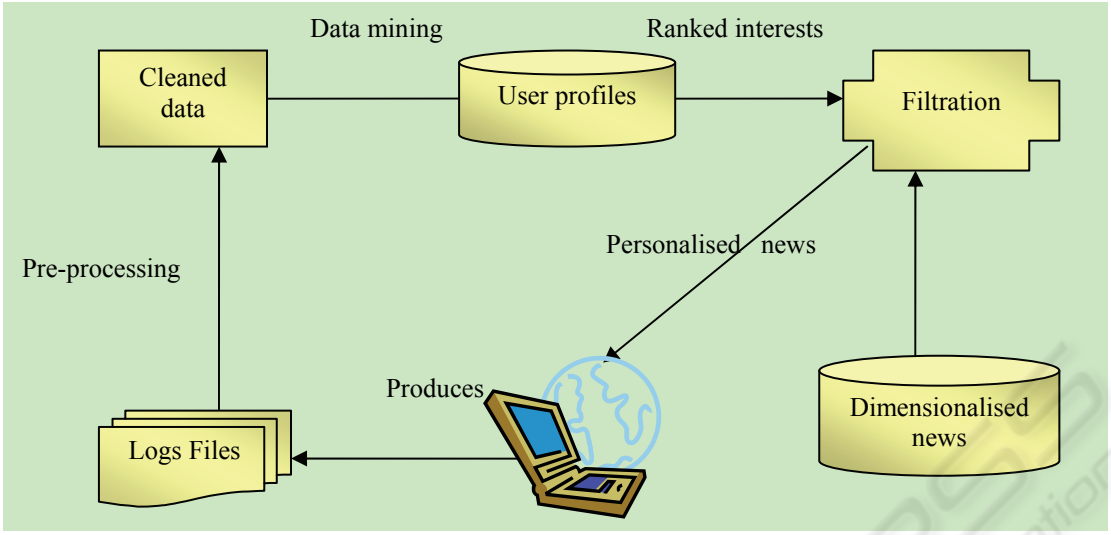


Figure 1: The general ANP architecture.

- ANP specified techniques identify sessions and users
- The resulting clean data is merged and analysed using data mining algorithms, particularly clustering. This discovers usage patterns used to construct user models/types.
- User profiles are created and a best match user model is assigned to each user.
- Finally ANP filters newly-arriving news items according to user profiles and provides a personalised output to each user.

Figure 1 shows how the ANP Architecture works.

Although the tasks involved in ANP are similar to those of other previous systems (Castellano et al., 2007; Grcar, 2004; Paliouras et al., 2006; Yang et al., 2006) the ANP uses additional techniques to provide better results. First, it achieves autonomy by monitoring users while they are browsing without the need to insert any information manually. Users are identified autonomously by their IP addresses, and the other tasks execute without any manual processes. Secondly, the system continuously monitors and analyses users' behaviour and produces improved profiles (to a finite depth). Thus ANP uses an adaptive personalisation cycle. Finally, the system should improve performance, as it does not require complex processing as part of browsing. It avoids the need for joins between huge data tables.

The result is that ANP provides a high degree personalisation without a significant performance overhead. A novel part of the ANP profiler is that a *Text Miner* is used to mine each news item and thereby derive its news category and location. Most existing systems maintain users' interests ranked by the category of news items they read. However, ANP adds another dimension, news location, allowing user interests to be ranked by both category and location. Figure 2 shows the difference between ranking interests in previous systems and in ANP.

| User interests | | |
|---|---|---|
| **Category** | **Location** | **Rank** |
| Politics | UK | 1 |
| Politics | Middle East | 5 |
| Sports | Brazil | 13 |
| Sports | UK | 4 |
| Weather | UK | 7 |
| . | . | . |
| . | . | . |

Interests Ranking (ANP)

| User interests | |
|---|---|
| **Category** | **Rank** |
| Politics | 2 |
| Economics | 1 |
| Sports | 7 |
| Family | 4 |
| Weather | 5 |
| . | . |

Adding location dimension

Interests ranking (Previous systems)

Figure 2: Interests ranking in previous systems vs. ANP.

## 4 ANP IMPLEMENTATION AND TESTING

SQL Server 2005 with Business Intelligence Studio was chosen to implement a prototype for ANP. The tool was used to perform most of the tasks involved

including connecting to data sources (log files), data preprocessing, data mining (pattern discovery), text mining and user profiling.

Figure 3 below shows the transformations involved in the data pre-processing phase. As shown, the process starts with the log flat file (row data) and ends with the clean data stored in a SQL Server destination. Raw data is first filtered by choosing the related fields/columns and records/rows. Then, some transformations are used to extract the news category and location as well as the session ID from each row of the log file. Finally, the resulted data is aggregated to analyse how many news items the user read and of which category and location. The resulted data is stored in a separate table in order to be analysed later.

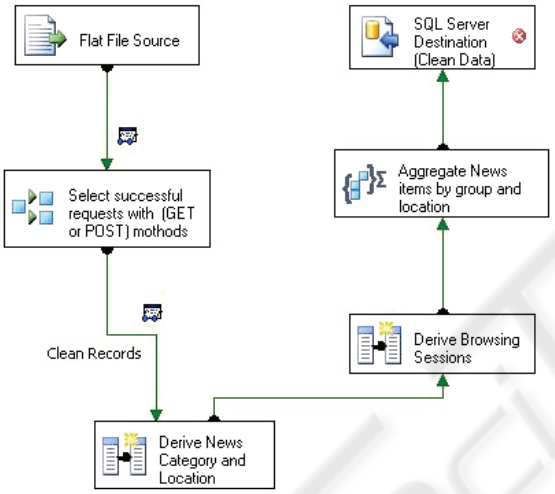


Figure 3: Transformations involved in the pre-processing phase.

For the data mining phase, Microsoft Clustering Algorithm is used to cluster users into groups based on their behaviour. The algorithm begins with identifying relationships between columns, and generating a series of clusters based on those relationships. After that, the algorithm calculates how well the clusters represent the groups of the points (users), and tries to redefine the groups to create clusters that represent the data in a better way. The algorithm iterates through this process until it cannot improve the results more by redefining the clusters. The result of the clustering process is illustrated by figure (4) below.

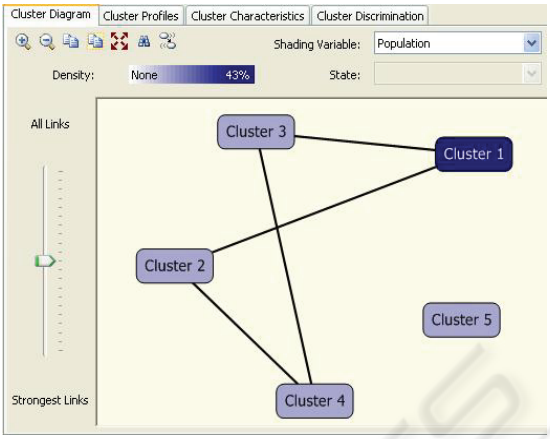


Figure 4: Results of applying Microsoft Clustering Algorithm.

Finally, the text mining task has not been implemented in this prototype since it requires an independent research project. Therefore, it has been left for further work.

To test the prototype, a fake news website was developed using ASP.net and a sample of 5 users browsed the site for a database of 1000 news items once daily for 5 minutes. The test lasted 6 days using Google and Yahoo for 3 days and then ANP for 3 days. All users were also given a questionnaire about news categories and locations in which they were interested. After using the test website for 3 days the ANP system was able to create a profile for each user. The average match between the results of the questionnaire and the automatically generated user profiles was 87.3% which suggests that the system recognised user interests with a reasonable level of accuracy even over such a short period of time. It is hoped that some improvement could be expected over a longer period. The next step was comparing the time the users spent in browsing relevant news items. Both Google and Yahoo News provide a poor degree of personalisation in addition to the lack of complete autonomy in delivering news. Initial results suggest that the ANP does provide some improvements. The average wasted time (calculated as a percentage of the 5 minutes daily browse time) for the 5 users was 2.2 minutes in Yahoo News, 2.75 in Google News but .75 minutes when using ANP.

These results are limited in that the users were looking at news in a range of locations as well as within the source country, which was the UK. They were therefore outside the broad spectrum of users expected on news sites who could be expected to look for news in their home country only. However, ANP was able to recognise these location

preferences and produce news from the users commonly requested locations. The browser was able to work autonomously in feeding news to users on the basis of the location and news category of the user's past browsing profile. It seems clear that adding location dimensions to the browsing history of each user is likely to produce improvements to news browser performance.

# 5 CONCLUSIONS

Personalisation has become an urgent need because users need to manage the massive data explosion in all information-based systems, particularly in web applications. Therefore, websites have started to offer personalisation services for their users, particularly in online news providing systems. In order to be efficient, a personalisation system needs to achieve four features: autonomy, adaptation to changes in users' behaviour, acceptable performance, and satisfactory matching to user requirements.

ANP is a prototype system designed to provide on-line-personalised news meeting the key features of personalisation outlined earlier, without affecting retrieval performance. The prototype provides a systematic method for managing personalisation by using web usage mining .

The results of implementing the prototype can be summarised in the followings:

- The system was able to connect to web log files and transform delimited values into a table of columns and rows.
- Logs raw data was successfully cleaned from noise in an intelligent way, with relatively noncomplex transformations. Non-required columns were not selected, where unrelated rows such as file headings, image, and unsuccessful records were filtered using several transformations.
- Users were identified by their IP addresses and browsing time was divided into sessions using certain transformations.
- After the data was preprocessed, it was summarised/aggregated according to user IP, news category and location, and session.
- The Microsoft clustering algorithm was applied successfully on the aggregated data, and resulted in a set of clusters. The clustering was efficient, and with the capabilities provided by SQL Server 2005, the results of clustering were refined further.

The developed prototype worked autonomously in performing the main system tasks, but not in all; because the system was not applied in a live scenario and there are still several issues to be addressed before this can be done. Furthermore, adaptation needs lots of log files and other resources in order to be implemented in a real context and this has been outside the scope of the immediate project.

# REFERENCES

Ardissono, L., et al., (2000b). Strategies for personalizing the access to news servers. [online] Stanford: AAAI Spring Symposium. Available from < www.di.unito.it/~liliana/ EC/aui00Giornale.ps.gz > [Accessed 12 August 2007].

Batista, P., and Silva, M. J., (2002). Mining Web Access Logs of an On-line. Malaga, Spain. 29-31 May 2002. eCTRL, 2002.

Castellano, G., et al., (2007). Log data preparation for mining web usage patterns. Proc. IADIS International Conference, Salamanca, Spain, 18-20 February 2007, Italy: University of Bari, 2007, 371-378.

Google, (2007). Google News. [online] Available from: <news.google.com> [Accessed 19 November 2007].

Grear, M. (2004). User Profiling: Web Usage Mining. Proc. The 7th International Multiconference Information Society IS, Ljubljana, Slovenia, 11-15 October 2004, IOS Press: Netherlands, 2004, 179-183.

Paliouras G., et al., (2006) PNS: Personalized Multi-source News Delivery. U.K., 9-11 October 2006. U.K.: Springer, 2006, 1152 – 1161.

Singh, S., et al., (2006). An Adaptive User Profile for Filtering News Based on a User Interest Hierarchy. In: Grove, Andrew, Eds. Proceedings 69th Annual Meeting of the American Society for Information Science and Technology (ASIST), Austin (US), 3-8 November 2006, 43, USA: Richard B. Hill, 2007.

Yahoo, (2007). Yahoo News. [online] Available from: <news.yahoo.com> [Accessed 20 November 2007].

Yang, Z., et al., (2006). An Effective System for Mining Web Log. Proc. of 8th Asia-Pacific Web Conference (APWeb'06), Harbin, China, 16-18January 2006, 40-52.