PersoNews: A Personalized News Reader Enhanced by Machine Learning and Semantic Filtering

E. Banos, I. Katakis, N. Bassiliades, G. Tsoumakas, I. Vlahavas

Department of Informatics, Aristotle University of Thessaloniki,
54124, Thessaloniki, Greece
{vmpanos, katak, nbassili, greg, vlahavas}@csd.auth.gr

Abstract. In this paper, we present a web-based, machine-learning enhanced news reader (PersoNews). The main advantages of PersoNews are the aggregation of many different news sources, machine learning filtering offering personalization not only per user but also for every feed a user is subscribed to, and finally the ability for every user to watch a more abstracted topic of interest by employing a simple form of semantic filtering through a taxonomy of topics.

1. Introduction

The explosive growth of the WWW has brought essential changes in everyday life. Maybe the most determining contribution was the boundless, instantaneous and costless offering of information. Recently, the rate of available information became gigantic, making the discrimination of useful information out of tons of worthless data a tedious task. This phenomenon is commonly named “Information Overload” and comprises a main issue impeding the user finding the needed information in time.

Machine Learning (ML) and especially Text Classification (TC) is a promising field that has the potential to contribute to the solution of the problem. In TC, a classifier is trained to separate interesting messages on behalf of the user. Much work has been done to this direction [8], but, unfortunately, not many applications were widely used.

We have implemented a system (PersoNews) in order to fill this gap. PersoNews is a web-based machine learning enhanced RSS reader. It utilizes an incremental Naïve Bayes classifier in order to filter uninteresting news for the user. The web application PersoNews has a twofold functionality. Firstly, it operates as a typical RSS reader, and secondly, the user can choose from a thematic ontology a *topic of interest.* The distinctiveness of PersoNews is that both the above functionalities are enhanced by the Machine Learning Framework described in [5].

In the rest of the paper, we cover related work on News Classification, in section 2; section 3 describes the ML framework chosen for the application and section 4 presents the PersoNews system. Our paper concludes with discussion and plans for future work.

2. Related Work

The problem of News Filtering is to effectively separate interesting news articles for the user from a large amount of documents. A typical application of this task would be to create a personalized on-line newspaper for each user. The role of machine learning in such problems was early recognized.

In [1], for example, a personalized on-line newspaper is created for every user, based on user feedback. The approach in that paper was to convert each article into a word/feature vector. Having the user profile also as a feature vector, all articles could be ranked according to their similarity with this vector.

In [3] a special purpose news browser for PalmOS-based PDAs is implemented. The authors use a Bayesian Classifier in order to calculate the probability that a specific article would be interesting for the user. An interesting part of this paper is the fact that the system does not take direct feedback by making the user evaluate every article. Instead, the news browser takes advantage of some other characteristics like total reading time, total number of lines, number of lines read by the user, and a constant denoting the user's average line reading time. All those metrics are utilized in order to automatically infer how interesting a particular user found a news article.

Billsus and Pazzani [2] implemented a Java Applet that uses Microsoft's Agent library to display an animated character, named News Dude, which reads news stories to the user. The system supports various feedback options like "interesting", "not interesting", "I already know this" and "tell me more about this". After an initial training phase, the user can ask the agent to compile a personalized news program.

Finally in [4] the user specifies his/her own categories of interest by entering keywords manually. These keywords are used in order to search for relevant articles in the world wide web. A classifier is used in order to filter uninteresting news. The classifier accepts feedback from the user who rates each article's relevancy.

Unfortunately, the aforementioned systems haven't been widely used, mainly because the problem was confronted as a personalization problem and not as an information overload problem, which appeared more recently. Moreover, most of these systems are constrained on specific sources of news articles. With the appearance of the RSS and OPML standards, it is far more straightforward to aggregate many different news sources. In addition, the user has his own personalized classifier per feed. That feature is crucial in dealing with information overload because an RSS feed can have many articles per day and the user will probably be interest for a small portion of those. Finally, in all the above systems, the user cannot declare general topics of interests like "databases" or "text classification".

In PersoNews, the user chooses a topic of interest from a thematic hierarchy. That functionality is of vital importance because a topic of interest can be covered more effectively by multiple news sources. Take as an example a user interested in a topic like "ComputerScience/DataBases". Interesting articles for this user could appear in many sources. For example, some articles may appear in a general Computer Science Journal, more relevant documents are going to be found in a more specialized source, like the proceedings of a database conference. At the same time, the user might be interested in commercial database management system like ORACLE. Therefore some interesting articles can be found in ORACLE's web site RSS feeds. In PersoNews all those sources are monitored under the same topic of interest, and additionally a classifier personalizes this monitoring for each user.

Contemporary popular Web RSS Readers like NewsGator (www.newsgator.com) and Google Reader (www.google.com/reader) can indeed aggregate and manage many RSS feeds but they luck of an abstracted thematic ontology and there is no machine learning filtering which will abate information overload.

3. Machine Learning Framework

In order to strengthen our system with an adaptive filter that will utilize user feedback a proper Machine Learning Classification system has to be chosen. Our problem (News Classification) can be classified as a Text Stream Classification problem with the (probable) occurrence of concept drift. Concept drift is the potential change of the target-class' concept in a classification problem. Therefore, we have the following requirements for our classifier:

1. An evidently good classifier for text categorization tasks.
2. An incremental classifier is required in order to constantly update knowledge when user sends feedback.
3. Because we intend to have a personal classifier for every user and for every feed or topic of interest the user subscribes, we needed a classifier with the minimum computational cost.
4. In a Text Streaming application there is no prior knowledge of the features/words that might appear and the use of a global vocabulary of hundreds of thousands of words is simply inefficient. Therefore, we need a classifier that has the ability to build dynamically the feature space as more documents/articles arrive. We call this space a "dynamic feature space".

Feature selection is of vital importance for text classification. Our two additional requirements for feature selection are:

1. It should be incremental and able to execute in a dynamic feature space.
2. An evidently good feature evaluation metric for text classification.

Classifiers that fulfil the first and third requirement are the Naïve Bayes (NB) classifier and Support Vector Machines (SVMs). We had to reject the use of SVM classifier due to complexity reasons and the fact that up to our knowledge there is no related work that describes how to execute SVMs in a dynamic feature space. Therefore SVMs do not meet requirements 3 and 4. The Naïve Bayes classifier on the other hand has shown good performance in text classification tasks [6] and is widely used in similar problems because of its simplicity and flexibility. Moreover, it can be easily incremental and as we discuss in previous work of ours [5] it can be straightforwardly converted into a feature based classifier, meaning that it can execute in a dynamic feature space. For the above reasons, the framework proposed in [5] is selected for use in PersoNews.

4. System Implementation

PersoNews is a web application which provides users with the ability to monitor a large set of web sites (RSS feeds) and receive notifications about new publications regarding topics of their interest. The system consists of three modules which function in parallel using a common database to store information. PersoNews system architecture is shown in Fig. 1.

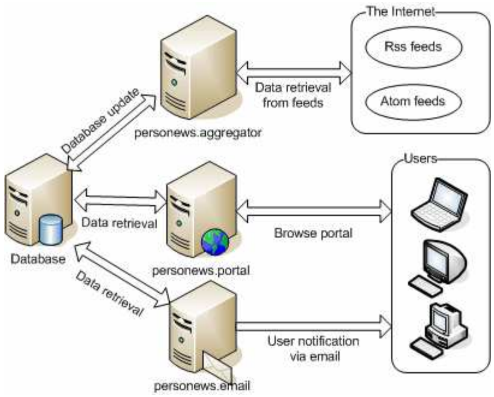


Fig. 1. PersoNews system architecture.

The main modules of PersoNews are:

Web site (PersoNews.portal). PersoNews.portal is responsible for the user interaction. Essentially, it is an open web application (news.csd.auth.gr) where everyone can register and have full access to PersoNews services using a web browser.

System update service (PersoNews.aggregator). PersoNews.aggregator is a server-side process which monitors RSS feeds in order to detect new publications and update PersoNews database. PersoNews.aggregator performs periodical polling of the news feed's URL in order to retrieve new publications which in turn are processed and stored in the database. Additionally, relevant topics are also updated.

Email notification service (PersoNews.email). PersoNews.email is responsible for notifying users by email about updates on feeds and topics they monitor. PersoNews.email is executed on a daily basis in order to check if there are any new publications in the feeds and topics monitored by each user. In that case, users are notified via email. PersoNews.email is fully customizable giving users the option to have email notifications for specific feeds and topics, change the email format or modify their email address.

4.1 Automatic News Classification

The most distinctive feature that adds up to the value of PersoNews is the integration of the Machine Learning framework discussed in the previous section We used Information Gain as a feature evaluation measure and Naive Bayes as a classifier. For the preprocessing step an implementation of Porter Stemmer [7] is used. Fig. 2 shows the preprocessing procedure for each publication.

While reading a publication from a feed or a topic, the user can perform some actions such as visit the publication's URL by clicking on its hyperlink. This action triggers a procedure that forces the PersoNews filter to update the classifier taking the specific document as an extra positive training example. Alternatively, the user can mark a publication as not interesting, forcing the PersoNews filter to utilize the specific publication as a negative training example and update the classifier. Users can also ignore new publications. In that case, the knowledge base is not updated at all.

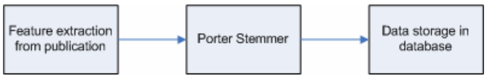


Fig. 2. PersoNews preprocessing steps.

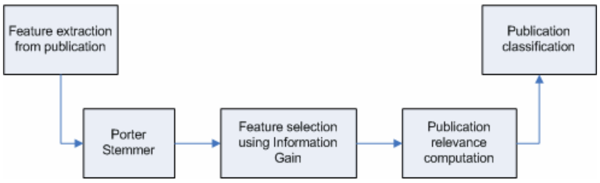

Feature extraction from publication
Publication classification
Porter Stemmer
Feature selection using Information Gain
Publication relevance computation

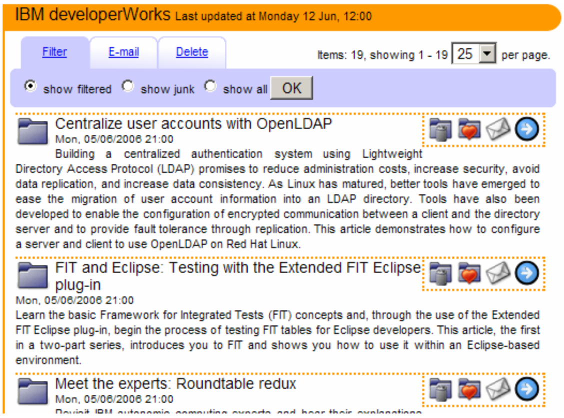
Fig. 3. PersoNews filter performs automatic classification of a new publication.

PersoNews filter training is performed incrementally, resulting in the creation of a knowledge base which includes various publication features as well as how much do they interest each user. As a result, when PersoNews.aggregator retrieves a new publication and extracts its features, it can decide in which extend it interests the user or not based on previous user feedback. Publications classified as interesting are displayed to users and are also sent to them via email while uninteresting publications are suppressed, thus reducing information overload. Fig. 3 shows the automatic classification of a new publication. It must be noted that each user has his own classifier for each subscribed feed and topic.

4.2 Feed Manipulation and Monitoring

Users can start monitoring feeds by selecting them from the list provided by the system or by entering their own feed URL. PersoNews also supports the OPML Protocol in order to batch import any number of feeds. It must be noted that due to the large number of feeds, they are organized into abstract categories according to their topics to enable better selection and browsing for users.

Clicking on a feed title allows the user to view its publications. Fig. 4 shows a list of publications in a feed. On the top right of each publication there are four icons which correspond to the available user actions (mark as junk, forward to a friend and visit URL).



IBM developerWorks Last updated at Monday 12 Jun, 12:00

Filter | E-mail | Delete

Items: 19, showing 1 - 19 25 per page.

● show filtered ○ show junk ○ show all OK

Centralize user accounts with OpenLDAP
Mon, 05/06/2006 21:00
Building a centralized authentication system using Lightweight Directory Access Protocol (LDAP) promises to reduce administration costs, increase security, avoid data replication, and increase data consistency. As Linux has matured, better tools have emerged to ease the migration of user account information into an LDAP directory. Tools have also been developed to enable the configuration of encrypted communication between a client and the directory server and to provide fault tolerance through replication. This article demonstrates how to configure a server and client to use OpenLDAP on Red Hat Linux.

FIT and Eclipse: Testing with the Extended FIT Eclipse plug-in
Mon, 05/06/2006 21:00
Learn the basic Framework for Integrated Tests (FIT) concepts and, through the use of the Extended FIT Eclipse plug-in, begin the process of testing FIT tables for Eclipse developers. This article, the first in a two-part series, introduces you to FIT and shows you how to use it within an Eclipse-based environment.

Meet the experts: Roundtable redux
Mon, 05/06/2006 21:00
Revisit IBM autonomic computing experts and hear their explanations

Fig. 4. Sample feed view.

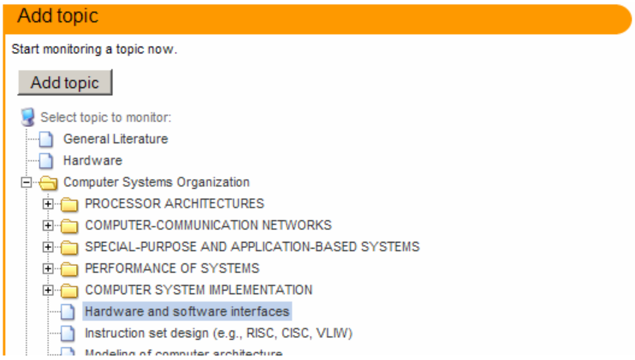
In case an article is marked by the user as not interesting, the document is forwarded to the classifier as a negative example in order to update its knowledge. When the user follows a URL, the system assumes that the user was interested in this publication and forwards the document to the classifier as an extra positive example. If the user follows the link and finds out that the article was not interesting, he/she can still mark the article as junk. In that case, the document is forwarded to the classifier as junk and we retrieve the previous positive example from the database. The user has also the ability to go into the “junk” folder and mark something as “not-junk” if he finds a misclassified article.

4.3 Semantic Filtering through Topic Manipulation and Monitoring

Except from monitoring specific feeds, users are able to monitor publications regarding a special topic of interest, such as “Database Management”, that belongs to the system's domain specific topic hierarchy, regardless the source feed of the publications. PersoNews has the ability to check all the feeds it is monitoring and locate new publications relevant to the topic. Currently, the topic selection list is a variant of the ACM Computing Classification System and it is organized in a tree structure featuring multiple levels of topic abstraction (Fig. 5). The topic hierarchy is implemented using an XML file to store topic descriptions as well as the associations between them. Notice that PersoNews is domain-independent, which means that it can operate under any topic hierarchy.

For each topic, the user can define one or more related keywords, which are essentially a set of words that act as an extra filter for new publications. If a publication from any feed contains any of these words then it is considered relevant to the topic. Initially, topic keywords derive from subtopics in the ACM topic hierarchy but users can also add their own custom keywords if they consider it appropriate (Fig. 6).

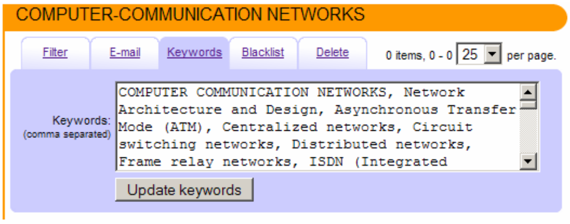
Under this operation, PersoNews employs a primitive form of semantic filtering, since each topic is accompanied by a number of user-defined keywords that supposedly describe the topic and can be considered as topic synonyms. Furthermore, the hierarchy of topics is also taken into account, since the keywords of all sub-topics are also considered to describe all their super-topics.



Start monitoring a topic now.

Select topic to monitor:
- General Literature
- Hardware
- Computer Systems Organization
  - PROCESSOR ARCHITECTURES
  - COMPUTER-COMMUNICATION NETWORKS
  - SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS
  - PERFORMANCE OF SYSTEMS
  - COMPUTER SYSTEM IMPLEMENTATION
  - Hardware and software interfaces
  - Instruction set design (e.g., RISC, CISC, VLIW)
  - Modeling of computer architecture

Fig. 5. Form used to start monitoring a topic by selecting it from the ACM topic hierarchy.



0 items, 0 - 0 25 per page.

Keywords: (comma separated) COMPUTER COMMUNICATION NETWORKS, Network Architecture and Design, Asynchronous Transfer Mode (ATM), Centralized networks, Circuit switching networks, Distributed networks, Frame relay networks, ISDN (Integrated

Fig. 6. Topic keywords.

As soon as PersoNews.aggregator retrieves new publications from feeds, it performs filtering in two steps in order to detect if there are any relevant items for each topic. Initially, it scans each new publication to check if it matches any of the keywords of every user's topics and in that case the classifier of that topic examines if the publication is interesting or not for the user. It is obvious that keywords act as an additional filter which performs a selection among all new publications. The ones which are selected from the keywords filter will have to transcend the PersoNews classifier as well.

When the user starts monitoring a topic, PersoNews searches for subtopics and includes them in the keywords of the selected topic. For example, when choosing to monitor the topic "Database Systems", PersoNews automatically aggregates every subtopic keywords such as "Database Concurrency", "Distributed databases", "Multimedia databases", "Object oriented databases".

MyTopics publications are displayed like MyFeed publications but, in addition, users can visit the source feed or start monitoring the source feed of the publication, or they can add the source feed to the topic blacklist (and therefore not receive publications from these sources anymore).

5. Conclusions and Future Work

Observing the concurrent growth of the World Wide Web and Information Overload we hope that systems like PersoNews will be widely used. Unfortunately the system was up until currently in beta version and thus, we did not encourage users to register until recently. Although an exhaustive evaluation is in our immediate plans, we had a crude estimation of the system's performance taken from a small amount of registered users. We had statistics showing that we had an average of 2.6% false negative rate (percentage of messages that the classifier marked as interesting but the users moved to the junk folder) and 5% false positive rate (percentage of messages that the classifier marked as junk but the users moved to the interesting folder), after a month of training. Although we do not claim statistical adequacy of this evaluation, we believe that these numbers are indeed encouraging.

Aside from an extensive evaluation of PersoNews, it is in our future plans to make the hierarchy offered by the system fully customizable, meaning that the user could add certain concepts in any level of the hierarchy. We also plan to investigate alternative machine learning frameworks which combine good scalability and performance. An essential important element we plan to add to our system is the aggregation of news sources that have no RSS feed available. For this purpose we investigate the potential of using Content Extraction techniques in order to extract text from simple html pages and create a corresponding RSS feed.

Acknowledgments. This work was partially supported by a PENED program (EPAN M.8.3.1, No. 03EΔ73), jointly founded by the European Union and the Greek Government (General Secretariat of Research and Technology).

References

[1] Bharat, K., Kamba, T., and Albers, M., *Personalized, interactive news on the web.* Multimedia Systems, 1998. **6**(5): p. 349-358.

[2] Billsus, D. and Pazzani, M. *A Hybrid User Model for News Story Classification*. in *Seventh International Conference on User Modeling*. 1999. Banff, Canada: Springer-Verlag.

[3] Carreira, R., et al. *Evaluating adaptive user profiles for news classification*. in *9th International Conference on Intelligent user Interface*. 2004. Funchal. Madeira, Portugal: ACM Press.

[4] Chan, C.-H., Sun, A., and Lim, E.-P. *Automated Online News Classification with Personalization*. in *4th International Conference of Asian Digital Library (ICADL2001)*. 2001. Bangalore, India.

[5] Katakis, I., Tsoumakas, G., and Vlahavas, I. *On the Utility of Incremental Feature Selection for the Classification of Textual Data Streams*. in *10th Panhellenic Conference on Informatics (PCI 2005)*. 2005. Volos, Greece.: Springer-Verlag.

[6] McCallum, A. and Nigam, K., *A Comparison of Event Models for Naive Bayes Text Classification*, in *AAAI-98 Workshop on Learning for Text Categorization*. 1998.

[7] Porter, M.F., *An algorithm for suffix stripping.* Program, 1980. **14**(3): p. 130-137.

[8] Sebastiani, F., *Machine Learning in Automated Text Categorization.* ACM Computing Surveys, 2002. **34**(1): p. 1-47.