A Personalized News Portal

Dung Dam Yang Wang
722007594 321004893
Texas A&M University
Department of Computer Science
and Engineering
{dungdt, wyang1989}@tamu.edu

ABSTRACT

With the numerous information generated everyday on the Internet, choosing news articles to read is not always an easy task for the online readers. They have to scan over many different titles to find an article that they are interested in. They may use the website's searching and filtering to narrow down the articles list but those functionalities are not always efficient and easy to use. Readers may have to repeat doing those stuffs from website to website just to find the articles they actually want to read. The whole process is time consuming and tedious because those websites are built for everyone, they are not personalized just for any specific user.

Observing that people are becoming more and more engaged with the social networks, we believe that it is possible to learn about their characteristics including hobbies, interests and concerns by discovering their social network profiles. By that, we are building a website that will gather the latest news from different sources and offer the users with social network account a personalized reading experience. We call our upcoming product a Personalized News Portal. In this report, we will present our observations of the contemporary news portals, our approach to exploit those observations in our product, the design and implementation of our product as well as some results and evaluations that we have till now.

1. INTRODUCTION

While reading newspapers is very popular today, people are still struggling to choose the favorite articles among millions of new titles created by thousands of news providers every day. What we are going to offer to the users is a single news portal where the users can view the new articles from many different sources.

Another problem to the readers is that, the news websites do not offer the personalization to any specific user. By personalization, we mean the interests and areas of concern of the users. The news articles often cover a wide range of topics, while the users actually want to read a small part of those. To bring the readers a personalized reading experience in our product, we need to learn about their characteristics, interests and concerns. We observe that by discovering the profiles of users on social networks, we can somewhat figure out their characteristics, as people are using and personalizing their own profiles every day. We focus on exploiting Facebook first, and our approach is to extract the users' interests and concerns by learning their Facebook data.

By solving the two problems above, we hope to offer the online readers a personalized news portal where they do not have to struggle to find the news. The next sections will show more details about the technical approach, results and evaluations of this product.

2. RELATED WORK

There are several news websites that offer the personalization of topics to be displayed with some effort from the readers.

The Google News [2] supports users to choose the news which they want to see from different sections. Users can decide how often they want to see articles in specific topics by configuring the frequency of articles in any specific topic, for example more often in Technology news and less often in Health. However they need to adjust these configurations manually and the options is also limited to only several big topics.

Prismatic [3] is another example of news portal where users can manually personalize the display by selecting the topics or publishers they are interested in. This website actually use machine learning on social networks to figure out what users should read. Users will need to log in with a social network account such as Facebook or Twitter and choose the topics and news providers that they are interested in, then the website will suggest them the articles by analyzing that selection and the users' social network data.

3. SYSTEM OVERVIEW

Two main features of our Personalized News Portal are gathering news from various sources and automatically personalizing the articles to be displayed based on analysis of reader's interests and concerns. Corresponding to these two features, the system backend contains two main components: a news collector and a news provider. One last component is a website, where the users can login with their Facebook account and view the news personalized for them. Figure 1 shows the overview architecture of the system.

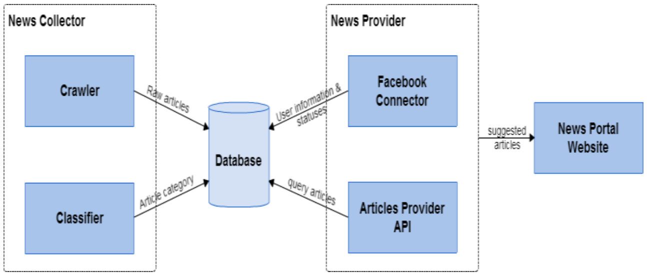


Figure 1 - System Overview

The news collector consists of a news crawler and a Naïve Bayes classifier. The main function of the news collector is to gather the news articles from several news providers, preprocess them, classify them and store them into a relational shared database. The crawler runs daily to download the latest articles from several predefined sources via the RSS news feed. From those, the crawler then extract the information including the title, content, image, published date and original categories of the articles, and insert extracted articles to database. The classifier in another process will add the articles to predefine categories. Next section will provide more details about the classifier.

The news provider is responsible for retrieving user information from Facebook when they log in, learning about users' interests and concerns and suggesting articles based on this learning. It includes a Facebook connector and an API for the website to get the articles.

The website provides a simple UI design, where the user can login with Facebook account and view the list of articles suggested by the system.

All the backend components (news collector and news providers) are written in Python using several packages including NLTK [4] for machine learning tasks and BeautifulSoup [5] for the information extraction in crawler. The database is MySQL. The website frontend is based on PHP. The frontend communicates with the backend via a REST API offered by the news provider.

4. METHODS

In our system, the categories of the news play an important role in the suggesting mechanism. For that reason, we need to put the articles into our predefined categories once they are downloaded and extracted by the crawler. Those categories will be displayed in the frontend for the readers to narrow down the topic of news they want to read. Also, category will be a heavy weighted feature of articles that the suggesting engine will base on to decide which articles to show to the readers.

4.1 Naïve Bayes Classification

For the classifier, we choose a probability approach with Naïve Bayes algorithm. Though this is a supervised learning algorithm, it needs a training dataset in order to train the classifier before it can classify the articles. We use the dataset Reuters-21578 containing the articles collected by Reuters for training and testing purpose. The dataset has more than 11,000 documents in training set and more than 4,000 documents in test set. The documents are already categorized with 115 labels in total. Though the labeled data is quite skewed, some categories have more than 1,000 documents, but some have less than 100, we have to merge categories into six groups: agriculture, business, industry, life, money, and politics. Each group contains several original categories so that the number of documents in these groups become more balanced. The documents are tokenized by spaces and punctuations, all stop words and words less than 3 characters are removed. All tokens are converted into lowercase. We use tf-idf to create feature vector for every document in the set. The documents will then be used to train the NLTK Naïve Bayes classifier. The same method of processing text is used to extract the feature vector from documents in the classifying process.

4.2 News Provider and Suggesting Mechanism

The news provider includes a Facebook API connector and news suggesting engine. We use Facebook PHP API to handle user login action. If the user has not already logged in, a window will prompt out to ask the authorization from the user. Then the user will be redirected back to our website after logging in with Facebook account.

The user's information will be stored in our MySQL database. This information includes user Facebook ID, hometown, education, interests, favorite teams and athletes, movies, working history and the latest 100 statuses. Currently we are trying to use these fields to represent the personal preference of this user.

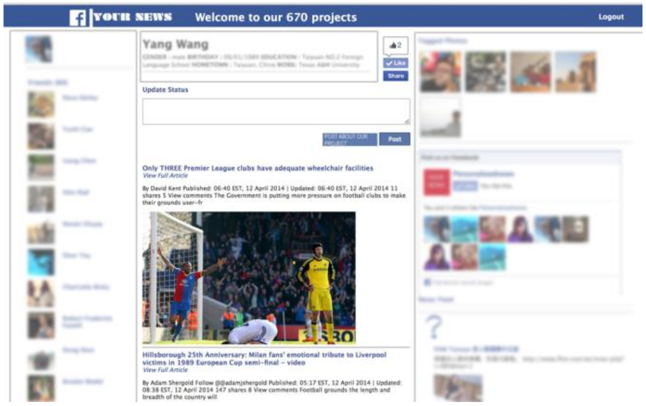
The suggesting engine is implemented in Python. It will get user's information and articles directly from database. These articles and user's information will be treated as feature vectors and the engine will calculate their tf-idf scores under inverted index system. Each article will be compared with the user vector by cosine similarity score to find the most suitable articles to suggest the user. The articles' score are increased if their category is the same as user's interests. For example, if the original rank is DOC1=0.89, DOC3=0.80, DOC2=0.77, DOC4=0.65. And now DOC4's category is sport and this user also likes sport category in their Facebook profile. We will increase the score of DOC4 by time 1.3. Then the new ranking will be DOC1, DOC4, DOC3, DOC2. The final score of some articles can be greater than 1.0. User's Facebook favorite categories and articles' categories are all map to our pre-defined category so we can make them comparable.

In the front-end, each news article contains the following fields: the title, summary of the content, source of article and an image. Users can click on the title to access the original article on provider website or they can also click on "View full article" to see the full content of the article on our website.

5. EVALUATION AND RESULTS

We run the crawler for a few days and retrieved about 5,000 latest titles from 3 main providers: Reuters, NYTimes and Dailymails.co.uk. The accuracy of the Naïve Bayes Classifier is around 80% when testing on the Reuters dataset. With the crawled dataset, we manually map the raw categories given by the providers with our categories and run the test. However, the classifier does not perform well on this crawled dataset, where the accuracy drops below 50%. One reason for this low performance is that the Reuters-21578 dataset is too old (from 1987) and it contains the many words that are not used much nowadays. The training corpus cannot cover well the real articles from Reuters, NYTimes and Dailymails. Another reason can be the way we empirically merge categories in Reuters-21578 into bigger groups does not reflect well the content of the documents. Some categories are merged into one group but they may have very different content, which leads to the diversity in the feature vectors of one group.

Result of the front-end. Figure 2 the screenshot of our website. When users view the recommended news, they can also post their status and view their news feed from Facebook at the same time. In order to achieve a quantitative evaluation of the suggesting engine, we plan to add the Thumb up/Thumb down button to each article and promote the website to get feedback from the real users. If the thumb up says users like the article, and thumb down says the opposite, we can measure the number of thumb up articles over total articles read by a user to get the percentage of correctly suggested articles. We can also dig into the thumb down articles to figure out what need to be improved in the suggesting mechanism.



Only THREE Premier League clubs have adequate wheelchair facilities
View Full Article

By David Kent Published: 06:40 EST, 12 April 2014 | Updated: 06:40 EST, 12 April 2014 11 shares 5 View comments The Government is putting more pressure on football clubs to make their grounds user-fr

Hillsborough 25th Anniversary: Milan fans' emotional tribute to Liverpool victims in 1989 European Cup semi-final – video
View Full Article

By Adam Shergold Follow @@adamjshergold Published: 05:17 EST, 12 April 2014 | Updated: 08:38 EST, 12 April 2014 147 shares 8 View comments Football grounds the length and breadth of the country will

Figure 2 - Website Screenshot

6. CONCLUSIONS

In summary, we have built successfully the personalized news portal with limited functionalities. Users can actually go to the website and login with their Facebook account to read the news. However, there are still many areas need to be improved, including the variety of the news and the performance of the machine learning algorithms. In the upcoming time, we want to increase the number of news providers appearing in the website to at least 20. That means we will need a better crawling strategy to overcome the Internet bandwidth problem. We also want to improve the accuracy of the classifier by applying stemming in the tokenizer and adding more data to the training dataset. Moreover, more information from Facebook profile can be exploited and the weight assigned to each feature can be different from user to user by learning algorithms to better represent the personalization. We believe that our efforts will help to bring a more convenient reading experience to the people in the future.

REFERENCES

[1] Ahmad Assaf, Aline Semart, and Raphael Troncy. SNARC – An Approach for Aggregating and Recommending Contextualized Social Content. In *The Semantic Web: ESWC 2013 Satelite Events*, pages 319-326, May 2013.

[2] Google News - https://news.google.com/

[3] Prismatic News - http://getprismatic.com/

[4] Natural Language Toolkit (NLTK) - http://www.nltk.org/api/nltk.classify.html

[5] Beautiful Soup 4 - http://www.crummy.com/software/BeautifulSoup/bs4/doc/

[6] Ana C. Cachopo. Improving Methods for Single-label Text Categorization. In *PhD Thesis at Technical University of Lisbon*. July 2007.

[7] Fabian Abel, Eelco Herder and Daniel Krause. Extraction of Professional Interests from Social Web profiles. *UMAP, 2001*.

[8] Ilknur Celik, Fabian Abel and Geert-Jan Houben. Learning Semantic Relationships between Entities in Twitter. In *ICWE, Springer (2001).*