

Automated Online News Classification with Personalization

CHEE-HONG CHAN AIXIN SUN EE-PENG LIM

Center for Advanced Information Systems, Nanyang Technological University
Nanyang Avenue, Singapore, 639798

Abstract

Classification of online news, in the past, has often been done manually. In our proposed Categorizor system, we have experimented an automated approach to classify online news using the Support Vector Machine (SVM). SVM has been shown to deliver good classification results when ample training documents are given. In our research, we have applied SVM to personalized classification of online news. In personalized classification, users can define their personalized categories using a few keywords. By constructing search queries using these keywords, Categorizor obtains both positive and negative training documents required for the construction of personalized classifiers. In this paper, we describe the preliminary version of Categorizor and present its system architecture.

1 Introduction

1.1 Motivation

Text classification is the process of assigning text documents to one or more predefined categories. This allows users to find desired information faster by searching only the relevant categories and not the entire information space. The importance of text classification is even more apparent when the information space is huge such as the World Wide Web. Examples of web classification systems include Yahoo! directory [15] and Google web directory [7]. However, such classification services are carried out by human experts, and they do not scale up well with the growth rate of web pages on the Internet. To automate the classification process, machine learning methods have been introduced. In a text classification method based on machine learning, classifiers are built (trained) with a set of training documents. The trained classifiers can therefore assign documents to their suitable categories.

Online news articles represent a type of web information that are frequently referenced. Currently, online news are provided by many dedicated newswires such as

Reuters [11] and PR Newswires [10]. It will be useful to gather news from these sources and classify them accordingly for ease reference. In this paper, we describe a working news classification system, named Categorizor [1], that performs automated online news classification. Categorizor adopts SVM classification method to classify news articles into categories. These categories can be either a set of predefined categories, i.e., general categories, or special categories defined by users themselves. The latter are also known as the personalized categories. With personalized categories, Categorizor allows users to quickly locate the desired news articles with minimum effort.

1.2 Related Work

Text classification is a well-studied problem. Several methods have been proposed and many of them can be directly applied to news classification as long as there exists a good set of training documents for each predefined category [17, 6, 14]. Nevertheless, when the categories (i.e., personalized categories) are defined on the fly and training documents are not readily available, the classification problem will become much more complex. On the other hand, personalized classification is a form of personalization and there are several existing ways to support personalization.

In the *collaborative filtering* approach, each user is associated with a user profile. When the user profiles of two users are similar, news articles that are interested in by one of them will be automatically recommended to the other [2].

In another personalization approach known as *content filtering*, one or more sets of features each representing a different interest domain (personalized category) of a user is derived. News articles are then recommended based on the semantic similarity with each set of features. In this approach, the interest domain of a user is very much independent of that of another user.

In the *subscription-based* personalization approach, a user can manually subscribe to a subset of a large number of pre-defined news categories. These pre-defined categories are usually static. News articles will be assigned to them when they are created. The news articles in the subscribed categories are then delivered to the user through e-mail or web browser. In other words, the subscription-based personalization approach is rather straightforward and does not require much classification efforts. Most of the web sites achieve news personalization by adopting the subscription approach, e.g. Newscon-online [9, 4].

Among the above three approaches, we have chosen to adopt content filtering approach to support personalized news classification in Categorizor system. The main difficulty in using this approach is that the training documents required for the generation of the

personalized classifiers cannot be easily obtained.

2 The Categorizor

2.1 General Features

For a start, we have designed Categorizor to classify news articles from the Channel News Asia [3] and these articles are mainly financial news. Categorizor offers two kinds of classification, namely, *general* classification and *personalized* classification. In general classification, we have adopted a fixed set of categories from the Reuters collection [12]. The Reuters collection was chosen because its categories are closely related to financial services and economics. Classifiers are trained for these categories using the corresponding Reuters news as training documents.

The unique feature of Categorizor is that it allows users to create and maintain their personalized categories. Users can create their personalized news category by specifying a few keywords associated with it. These keywords are known as the category profile for the newly created category. There is no restriction on the number of personalized categories for each user. To build the classifier for a personalized news category, a number of training documents (news articles) have to be obtained. Instead of getting the user to perform the time-consuming task of selecting and uploading the training documents, we construct a query to the Yahoo News Search Engine [16] using the category profile, i.e., the keywords. The training documents are then selected from the most highly ranked resultant news articles from Yahoo news.

To cope with evolving user interests and to further improve the effectiveness of classifiers for personalized categories, our personalized classifiers are defined such that they can be retrained upon user request. When user read the classified news articles, user is able to give feedbacks on the news, that is, whether the news article is correctly classified to the personalized category. These news articles carrying feedbacks will be later used as training documents for retraining the personalized classifiers.

2.2 Architecture

The architecture of Categorizor is shown in Figure 1. The main architecture consists of six modules, that is, the *Pre-processing*, *Presentation*, *Storage*, *SVM Classifier*, *User Registration* and *Webpage Retrieval* modules.

The *Webpage Retrieval* module employs the web page crawler to download the online news articles from the Channel News Asia web site. The *Pre-processing* module consists

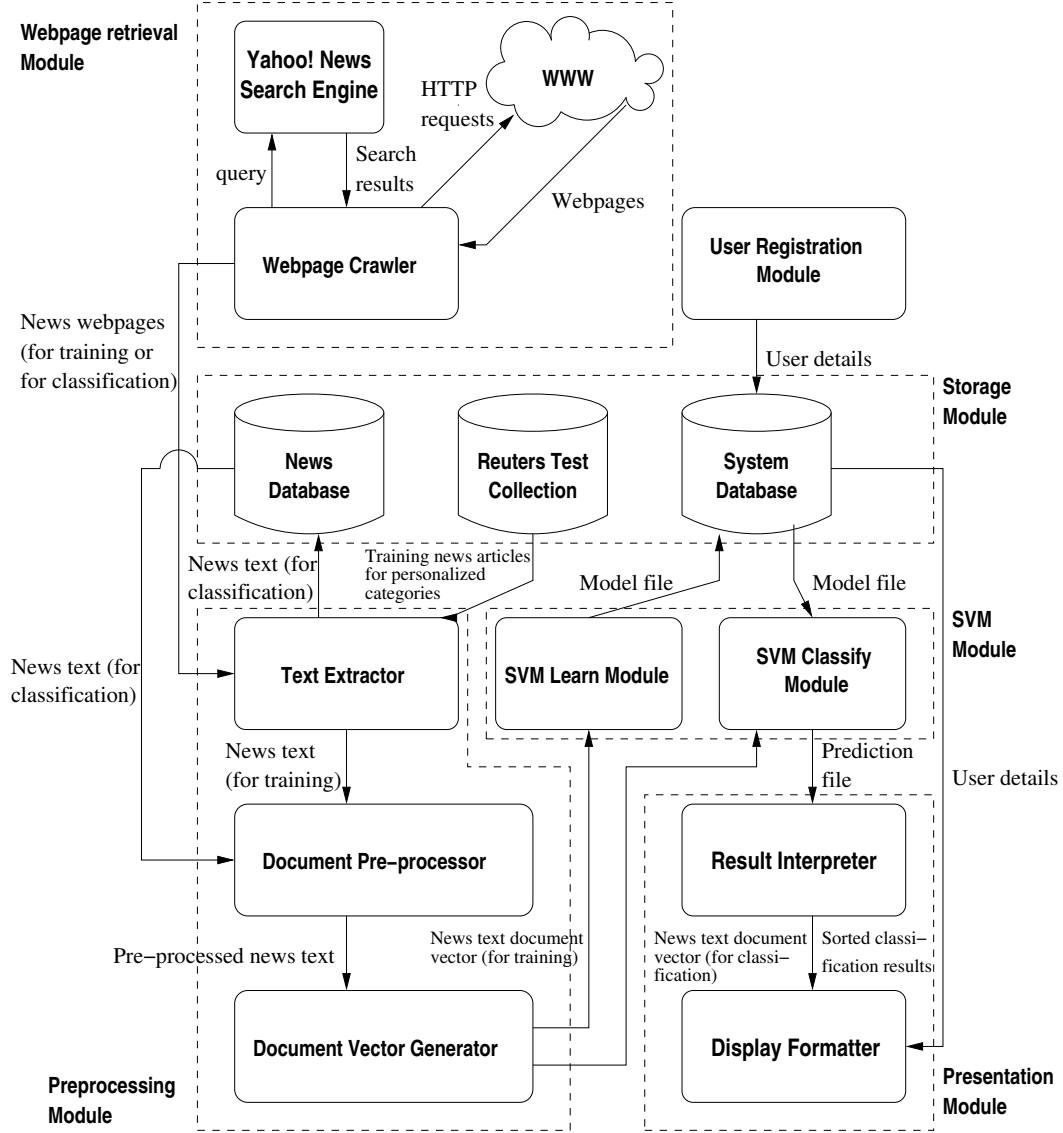


Figure 1: Architecture overview of the Categorizor

of the Text Extractor, Document Pre-Processor and the Document Vector Generator. The Text Extractor extracts the news text from the downloaded news web pages. The extracted news text is stored in the News Database. The Document Pre-processor performs stop-word removal and word stemming on the extracted text. After pre-processing, document vectors are generated by the Document Vector Generator using the well known $tf \times idf$ scheme [13]. To cater for documents with varying length, the document vectors are normalized to unit length.

There are three information repositories in the system. The *News Database* stores the attributes of the news articles downloaded from the online news web sites for both training (in the case of personalized classification) and classification. The attributes to be

stored include the downloading date, the URL and the news text. The *System Database* holds information about users and their personalized categories.

The *SVM Classifier* is a binary classifier which consists of the SVM Learn Module and the SVM Classify Module. The SVM Learn Module trains the classifier of a category (general or personalized) and produces a model file. Given the model file, the SVM Classify Module performs classification on a given set of documents (represented by their document vectors). In our prototype system, the *SVM^{light}* package developed by Joachim is used [8].

The *Presentation* module sorts the classification results from SVM classifier according to the score values returned by the SVM Classify Module. The *User Registration* module is responsible for the management of user information and their personalized categories.

3 Classification Process

Categorizor performs two kinds of classification as mentioned in Section 2.1. The two kinds of classification are performed in different ways. The detailed classification process are described in this section.

3.1 General Classification

In general classification, all the categories are taken from the Reuters-21578 text collection. Only 10 general categories are currently supported by the Categorizor and users can select any general categories for viewing as shown in Figure 2. We build a SVM classifier for each of the 10 categories as each SVM classifier is capable of giving a binary decision given an input document. The steps of training and using a SVM classifier are as follows:

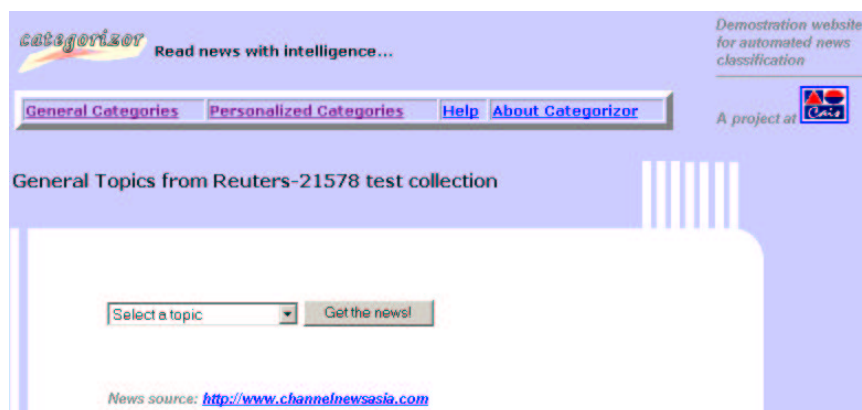


Figure 2: Selection of general categories

1. The SVM classifier is trained with the training documents from the Reuters-21578 text collection. The positive documents are the ones that belong to the category and equal number of negative training documents are randomly selected from the rest of categories. After training, the output of the SVM classifier (i.e. the model files) are stored in the System Database.
2. The news articles are downloaded daily from the source website, (i.e. the Channel News Asia news) and their text are extracted from the news bodies by the Text Extractor and then stored in the News Database. The text are referred as documents in the later process.
3. When the user requests for the news from category C_i , the most recently downloaded documents are retrieved from the News Database. Their document vectors are generated by the Document Pre-processor and Document Vector Generator.
4. The model file for category C_i is retrieved from the System Database and the corresponding SVM classifier will start classifying the document vectors.
5. The classification results are sorted according to the score values assigned by the SVM classifier and displayed in the resultant web page as shown in Figure 3. In the resultant web page, we use 5-point ranking to identify the relevance of the the news articles to the category.



Figure 3: Results returned for general categories

3.2 Personalized Classification

In personalized classification, the personalized categories are defined by users and each category is described by a few keywords. The classification steps are as follows:

1. The user first registers his/her user name and password with the Categorizor.
2. The user defines his/her personalized categories by providing category names and a set keywords for each personalized category that describe the content of the category.
3. To obtain the training news articles (documents) for each personalized category, the keywords are submitted to the Yahoo! news search engine and the news articles originally from Reuters returned by Yahoo! are used as the positive training documents. The negative training documents are obtained by conducting an inverse keyword search on the Yahoo news search engine. The inverse keyword search can be easily achieved by adding a “-” operator before the keyword. For example, the Yahoo! news search engine will return all the news that *do not* contain keywords “market” if “-market” is given as the search query.



Figure 4: Entering keywords to define a new personalized category

4. All the positive and negative training news articles are submitted to the Text Extractor and the document vectors are generated with the Document Vector Generator.
5. A SVM classifier is constructed by the SVM learn module for the each newly constructed personalized category. The learning process utilizes both the positive and negative training document vectors. The generated classifier is stored as a model file within the System Database.
6. When a user requests for news under his/her personalized category C_j , the recently

downloaded news from News Database are retrieved and their document vectors are generated by the Document Vector Generator.



Figure 5: Classifier re-training

7. Both the document vectors and the model file for C_j are passed to the SVM Classify module and the classification results sorted by score values are displayed in HTML format. In the resultant web page, a “Relevant?” checkbox is associated with each news entry, as shown in Figure 5, to allow feedback from the user.

3.3 Re-training of the Classifier

In order to strengthen the personalization aspect of Categorizor, it is designed to accept feedback from the user. The user, while reading news from a category can indicate if the content of the news article is relevant for the particular category by checking the “Relevant?” box. When the “Update classifier with selected document(s)” button is clicked, the corresponding classifier will be re-trained with a new training set that includes the feedbacked documents. In this way, users can constantly refine the training sets for their personalized categories with better accuracy. At present, we have not evaluated the

effect of re-training in personalized classification. Experiments to evaluate the different ways of training will be covered in the future research.

4 Conclusion

We have designed and implemented a preliminary version of news classification system based on the SVM classification method. The system is capable of both general classification and personalized classification. Our preliminary experiments, not reported in this paper, have shown that our system works well for the general classification while there are rooms for improvement for the personalized classification.

As Categorizor is still in its development and enhancement stage, much work need to be done to make it a full-fledge news classification system, particularly the personalized classification feature. Firstly, we need to enhance the Categorizor with a complete set of general categories. Due to the unavailability of a generic extraction software for extracting the desired news text from HTML web pages, Categorizor is currently restricted to classifying news articles from Channel News Asia only. A complete version of the Categorizor will have to incorporate an extraction facility that allows users to specify the sources of news articles. We are currently conducting experiments to improve the performance of personalized classification. For example, we are exploring the use of hierarchy in personalized classification as it has been reported that hierarchical classification gains better performance than the flat classification [5].

References

- [1] Categorizor, <http://www.cais.ntu.edu.sg:8000/~cheehong/servlet/categorizor>.
- [2] R. J. Chen, M. Nathalie, and W. Shawn. Collaborative information agents on the world wide web. In *Proceedings of the third ACM Conference on Digital libraries*, pages 279–280, 1998.
- [3] Channel News Asia, <http://www.channelnewsasia.com>.
- [4] R. Däßler, K. Schirmer, and G. Neher. Business news in 3 dimension, 1998. <http://fabdp.fh-potsdam.de/infoviz/paper/ieee98.pdf>.
- [5] S. Dumais and H. Chen. Hierarchical classification of Web content. In *Proceedings of the 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 256–263, Athens, GR, 2000. ACM Press, New York, US.

- [6] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, pages 148–155, 1998.
- [7] Google Web Directory, <http://www.google.com/dirhp>.
- [8] T. Joachims. *SVM^{light}*, an implementation of Support Vector Machines (SVMs) in C. http://ais.gmd.de/~thorsten/svm_light/.
- [9] Newscan-Online, <http://www.newscan-online.de/newscan/index.html>.
- [10] Pr newswires, <http://www.journalismnet.com/pr/prwires.htm>.
- [11] Reuters, <http://www.reuters.com>.
- [12] Reuters-21578 text categorization test collection, <http://www.research.att.com/~lewis/reuters21578.html>.
- [13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
- [14] F. Sebastiani. Machine learning in automated text categorisation: a survey. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell’Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999. Revised version, 2001.
- [15] Yahoo! Directory, <http://www.yahoo.com>.
- [16] Yahoo! News, <http://news.yahoo.com>.
- [17] Y. Yang and X. Liu. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkley, August 1999.