TV-Show Retrieval and Classification

Cataldo Musto[1], Fedelucio Narducci[1], Pasquale Lops[1],
Giovanni Semeraro[1], Marco de Gemmis[1],
Mauro Barbieri[2], Jan Korst[2], Verus Pronk[2], and Ramon Clout[2]

[1] Department of Computer Science, University of Bari "A. Moro", Italy,
{cataldomusto,narducci,lops,semeraro,degemmis}@di.uniba.it
[2] Philips Research, Eindhoven, The Netherlands,
{mauro.barbieri,jan.korst,verus.pronk,ramon.clout}@philips.com

Abstract. Recommender systems are popular tools to aid users in finding interesting and relevant TV shows and other digital video assets, based on implicitly defined user preferences. In this context, a common assumption is that user preferences can be specified by program types (such as documentary, sports), and that an asset can be labeled by one or more program types, thus allowing an initial coarse preselection of potentially interesting assets. Furthermore each asset has a short textual description, which allows us to investigate whether it is possible to automatically label assets with program type labels. We compare the Vector Space Model (VSM) with more recent approaches to text classification, such as Logistic Regression (LR) and Random Indexing (RI) on a large collection of TV-show descriptions. The experimental results show that LR is the best approach, but RI outperforms VSM under particular conditions.

Keywords: Vector Space Model, Random Indexing, Logistic Regression

1 Introduction

Automatic TV recommendations have been explored extensively in the literature where most papers assume that the set of items for recommendations is of moderate size. Most approaches are not directly applicable to web video repositories (such as YouTube) whose item sets are orders of magnitude larger. To provide personalized recommendations for digital assets on the web and TV, a possible approach is to match the assets' textual descriptions to personal preferences of users. It is common practice to classify TV shows by labeling them with one or more *program type* labels. It may also be assumed that user preferences can be coarsely expressed in terms of program types [2]. In this paper, we assume that each asset has a short textual description and we investigate (a) how well that description can be automatically mapped to a program type and (b) which machine learning algorithms are best suited for the above mentioned classification task. To this end, we have extensively tested algorithms using a large collection of TV-show descriptions which calls for the adoption of simple and scalable retrieval models. A text classification algorithm based on the Vector Space Model

(VSM) might be a good solution, provided that effective dimensionality reduction techniques are integrated, such as Random Indexing (RI) [3]. As regards classification algorithms, we opted for Logistic Regression (LR), since it is generally considered as accurate as Support Vector Machines, with the advantage of yielding a probability model [4].

This research is carried out in the context of a joint project with APRICO Solutions[3], a software company and part of Philips Electronics. APRICO Solutions develops video recommender and targeting technology, primarily for the broadcast and internet industries. Further details are available in [1].

2 TV-show Classification and Retrieval

The two problems we focus upon can be defined as follows:
TV-show classification: given a program description s and a set P of program types, choose a program type $p \in P$ that best matches the program description. Each TV show has exactly one label assigned to it.
TV-show retrieval: given a set S of TV-show descriptions and a program type $p \in P$, return a ranked list of k TV-show descriptions from S that best match program type p.

Three approaches for the TV-show classification and TV-show retrieval tasks have been investigated. We compare VSM with LR and RI. For both tasks, TV-show textual descriptions have been preprocessed for obtaining bag-of-words representations (BOW).

2.1 TV-SHOW CLASSIFICATION

Vector Space Model Given a set of documents (*corpus*), each document is represented as a point in a *n-dimensional* vector space (n is the cardinality of the vocabulary). Formally, each document is represented as a vector $\boldsymbol{d} = (w_1, \ldots, w_n)$ where w_i is the TFIDF score of the feature i. A vector space representation of each program type is obtained by summing the vectors of TV shows belonging to that program type. Thus, given a TV show s to be classified, its program type is given by the program type vector with the highest cosine similarity to s. VSM has some important limitations: it is not incremental and it does not model semantics.
Random Indexing. RI is a scalable and incremental dimensionality reduction technique. It belongs to the class of *distributional models*, which state that the meaning of a word can be inferred by analyzing its use (*distribution*) within a corpus of textual data. Random Indexing for TV-show classification follows the same steps as for VSM: a prototype vector is built for each program type and the cosine similarity between a TV-show and each program type is computed. Unlike VSM, these steps are performed on the reduced vector space obtained as output of the RI algorithm (500, 700 dimensions).

[3] www.aprico.tv

Logistic Regression. LR is a supervised learning algorithm based on a generalized linear model. In this work we exploited the implementation provided in LIBLINEAR[4]. Given a TV show, we compute the probability of each program type by exploiting the logistic functions learned for each class. The TV-show program type is determined by the highest probability.

2.2 TV-SHOW RETRIEVAL

For the TV-show retrieval task, we exploited only LR and RI, since they achieved the best performance for most classes in the classification task.
Random Indexing. As in the classification task, the vector space is reduced through the RI algorithm. Given a prototype vector built for each program type, the cosine similarity with all TV shows is computed in order to get the list of the best matching TV-show descriptions for a specific program type.
Logistic Regression. The probability that a TV show belongs to a specific program type is computed for the retrieval task as well. In this task, given a program type p, the TV shows are ranked based on their probability to belong to p and are returned in a ranked list.

Program Type	VSM	RI		LR
		500	700	
miscellaneous	0.11	**0.37**	**0.35**	**0.26**
movies	0.76	0.35	0.40	**0.83**
short movies	0.35	**0.95**	**0.95**	**0.75**
tv series	0.74	0.47	0.58	**0.87**
sport	0.90	**0.90**	**0.91**	**0.96**
show	0.65	0.48	0.48	**0.85**
events	0.63	**0.72**	**0.74**	**0.86**
documentary	0.63	0.23	0.24	**0.72**
reportage	0.57	0.41	0.43	**0.75**
report	0.15	**0.32**	**0.30**	**0.43**
magazine	0.64	0.39	0.36	**0.81**
news	0.54	**0.84**	**0.83**	**0.82**
videoclip	0.79	**0.94**	**0.92**	**0.83**
advertising	0.94	**0.99**	**0.99**	**0.98**
music	0.81	**0.83**	**0.81**	**0.84**

Fig. 1. Accuracy of VSM, RI, and LR for the classification task.

Alg	Dim	Precision@n%					
		5%	10%	25%	50%	75%	100%
RI	500	0.58	0.52	0.46	0.41	0.38	0.36
RI	1000	0.58	0.53	0.47	0.42	0.38	0.36
RI	1500	0.57	0.53	0.47	0.41	0.39	0.35
RI	2000	0.57	0.53	0.46	0.41	0.34	0.35
LR		**0.92**	**0.90**	**0.88**	**0.86**	**0.82**	**0.75**

Fig. 2. $P@n\%$ of RI, and LR for the retrieval task.

3 Experimental Evaluation

The goal of the experimental evaluation is to measure the effectiveness of the VSM, RI, and LR models in the retrieval and classification tasks. The experiment

[4] `www.csie.ntu.edu.tw/~cjlin/liblinear/`

has been carried out through a *k-fold cross validation* (k=10), on a dataset composed of 133,579 TV shows broadcast from a set of 47 channels in the German language. The textual descriptions are the input to the learning process and are represented by bag of words. Stemming and stop-words elimination are performed on the text. For the *classification* task we used the Accuracy as metric: it is calculated as the ratio between the TV shows correctly classified and the total number of TV shows classified. For the *retrieval* task we used the Precision@n%: it is calculated as the ratio between the TV shows correctly classified and the $n\%$ of the Test Set. VSM, LR, and RI (using different vector space dimensions) have been compared.

Classification task. Figure 1 reports accuracy values of VSM, LR and RI. The configurations that overcome the baseline (VSM) are in bold. For some classes the dimensionality reduction technique deteriorated the performance of the classifier. However for most classes, RI outperformed VSM, even though the reduction of the vector space dimension is considerable. Furthermore, the LR algorithm obtained the best accuracy. The best improvement achieved compared to the VSM model is almost 20%.

Retrieval task. In general the different space dimensions for random indexing do not affect the retrieval accuracy of the retrieval model (see Figure 2). Also for this task LR achieved better results compared to RI. The accuracy of the model decreases when the size of the retrieved list increases. This was expected because less relevant shows for each program type are in the tail of the list.

4 Conclusions and Future Work

The best performing approach for the classification task was LR. Despite the fact that this approach already showed to be effective in text classification in the literature, results achieved in this specific scenario were not obvious, since TV shows have very short textual descriptions and only few training examples were available for many classes. RI demonstrated a good performance in TV-show classification for the classes with a small number of instances in the training set. In the retrieval task LR outperforms the other approaches as well. In the future we will work in a recommendation scenario in order to re-rank the retrieved list of TV shows according to the user preferences.

References

1. C. Musto and F. Narducci. Tv-show retrieval and classification. Technical report, Philips Research, High Tech Campus, Eindhoven, The Netherlands, July 2011.
2. V. Pronk, J. Korst, M. Barbieri, and A. Proidl. Personal television channels: simply zapping through your pvr content. In *Proceedings of the 1st International Workshop on Recommendation-based Industrial Applications*, RecSys '09, 2009.
3. M. Sahlgren. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop, TKE 2005*, 2005.
4. T. Zhang and F. J. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4:5–31, 2000.