Knowledge-Based Systems 76 (2015) 53–66

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Semi-supervised evolutionary ensembles for Web video categorization ☆

Amjad Mahmood, Tianrui Li *, Yan Yang, Hongjun Wang, Mehtab Afzal

School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China

ARTICLE INFO

Article history:
Received 9 May 2014
Received in revised form 25 November 2014
Accepted 28 November 2014
Available online 16 December 2014

Keywords:
Genetic algorithm
Semantic similarity
Clustering ensemble
Social media mining
Video categorization

ABSTRACT

Evolutionary Algorithms (EA) have been developing rapidly as a powerful and general learning approach which has been used successfully to find a reasonable solution for data mining and knowledge discovery. Genetic algorithm (GA) is a kind of mainstream EA paradigm with a purpose of developing solutions for optimization problems. Clustering ensembles have emerged as an outstanding algorithm in machine learning to leverage the consensus across multiple clustering solutions and combines their predictions into a single solution with improved robustness, stability and accuracy. Multimedia advancement and popularity of the social Web has collectively provided an easy way to generate bulk of videos. Categorization of such Web videos has become a hot research challenge. In this paper, we propose a Semi-supervised Evolutionary Ensemble (SS-EE) framework for social media mining, e.g., Web Video Categorization (WVC), using their low cost textual features, intrinsic relations and extrinsic Web support. The contributions of this research work are as follows. First, we extend the traditional Vector Space Model (VSM) to Semantic VSM (S-VSM) by considering the semantic similarity between the feature terms using Normalized Google Distance (NGD) approach. Second, we define a new distance measure, Triangular Similarity (TrS) between two Textual Feature Vectors (TFV) based on the frequencies of most relevant terms in each category. Third, we iterate the clustering ensemble process with the help of GA guided by a new measure, Pre-Paired Percentage (PPP), to be used as the fitness function during the genetic cycle. Fourth, in the key steps of the GA, crossover and mutation genetic operators, we define them by an intelligent mechanism of clustering ensemble. Fifth, in order to terminate the genetic cycle, we define another new measure, Clustering Quality (Cq), based on similarity matrix and clustering labels. Experiments on real world social-Web data (YouTube) have been performed to validate the SS-EE framework.

1. Introduction

Automatic Web Video Categorization (WVC) provides an auspicious way to identify the categories of Web videos such as film, gaming, and education. It performs a promising role in effective and efficient browsing as well as retrieving the large corpus of Web videos. However, the challenges started from abundant data diversity within a category, deficiency in precisely labeled training data, and atrociousness of video quality, make the analysis of diverse Web videos (such as YouTube [1]) a challenging task. A lot of research work has been carried out on this issue by utilizing visual, textual and audio features individually or with different combinations to train models for Web video classification [2]. These methods depend mostly on building models through machine learning techniques (e.g., SVM, HMM, GMM) to map visual low-level features to the high-level semantics. Due to inadequate results of high-level concept detection methods [3] and expensive feature extraction, the content based categorization could not achieve the required results. Therefore, using textual information associated with the Web videos may become a feasible approach for the categorization purpose.

Basic text oriented mining approaches are quite ineffective to well categorize the YouTube videos due to inadequate textual details, e.g., title and tag which are brief phrases with noisy, incomplete, and ambiguous terms, whereas the description consists of a few brief sentences but with limited keywords. Besides supervised learning (classification), where some amount of pre-labeled data is used for learning purposes, unsupervised learning (clustering) has a unique role in data mining research. Clustering ensemble is an advanced approach in this area. Genetic algorithms (GAs) are well-known for being highly effective in optimization problems by formulating the search method that can be used both for solving problems and modeling evolutionary systems [4]. GA has proved a

☆ This is an extended version of the paper presented at the ADMA2013, Hangzhou, China.

* Corresponding author.
E-mail addresses: amjad.pu@gmail.com (A. Mahmood), trli@swjtu.edu.cn (T. Li), yyang@swjtu.edu.cn (Y. Yang), wanghongjun@swjtu.edu.cn (H. Wang), mehtabafzal@gmail.com (M. Afzal).

http://dx.doi.org/10.1016/j.knosys.2014.11.030
0950-7051/

prominent role in finding consensus cluster partitions during clustering ensemble.

Based on information distance and descriptive complexity, a method was developed to calculate the similarity between words from the WWW using Google page counts [5]. The popularity of the Internet and ease of access has motivated millions of users to input billions of words to create trillions of Web pages of, on average, low quality contents. Normalized Google Distance (NGD) is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords [6].

Similarity between two Textual Feature Vectors (TFVs) is an important step in text mining. Corresponding feature matching is a traditional approach. A large corpus of data can provide frequencies of feature terms which ultimately lead to term weights for a particular category. Two TFVs, instead of comparing with each other directly, can also be compared with term weights of Category Feature Vector (CFV) to represent their relation. The identical behavior with respect to such relations can lead to calculation of the similarity measure between two TFVs. This additional similarity measure can provide a reasonable support to clustering ensemble process in the genetic cycle.

A general approach to evaluate the clustering results is the use of validated datasets, but the availability of such datasets is not guaranteed in all situations. In such situations, we propose a new measure, Clustering Quality (Cq), that can be used to evaluate the clustering ensemble outcome. Although there are many methods available for measuring the quality of clusters [7,8], they depend upon the information available in particular situations. At present, we have the similarity matrix and label information of videos. Based on such information, we employ the Cq measure for a clustering solution to evaluate and compare two consecutive solutions. This measure can provide a right and timely direction at the termination stage in the genetic cycle to either terminate the cycle or continue for next iteration.

In our previous work, we proposed a Semi-supervised Cluster-based Similarity Partitioning Algorithm (SS-CSPA) [9] and a Semi-supervised Cluster-based Similarity Partitioning Algorithm evolved by GA (SS-CSPA-GA) [10] to cluster the videos containing textual data and related video information. We extend this work and aim to deal with the categorization problem of Web videos, by an evolutionary learning process of GA using their textual data based on the semi-supervised clustering ensemble approach. The paper then argues that, in order to improve the quality of base clusters, extension of the traditional Vector Space Model (VSM) using NGD support from Google is more favorable approach. The paper also proposes the Twin Behaviour Model (TBM) by introducing the Triangular Similarity (TrS) measure using a large corpus of data. This paper also emphasizes the appropriate translation of related video information in terms of Must-Link (ML) constraints by using our proposed Must-link Mesh Model (MMM), which ultimately are used to define Pre-Paired Percentage (PPP) measure. In addition, instead of using validated dataset accuracy for termination of the genetic cycle, a Cq measure is proposed in terms of the similarity matrix and resulting labels.

The rest of the paper is organized as follows. In Section 2, a brief survey of related work is described. Section 3 introduces the preliminary definitions that are used in our proposed framework. Section 4 demonstrates the proposed framework for WVC. Section 5 shows the experimental details and evaluation of results. Concluding remarks and future work are stated in Section 6.

2. Related work

Multimedia advancement and popularity of social media has provided a bulk amount of videos making the selection criteria more complicated for a user in finding his required video. This motivates the researchers to design classification algorithms to manage all these videos into their respective categories, usually with a semantic label attached to each, i.e., News, Sports, Comedy, etc. The early goals in this direction were achieved by [11], and then further improved by using classifiers from open Web resources for large-scale video classification [12]. Further contributions towards video genre classification were proposed by [13] in which object motions, audio, color statistics, cut detections and the camera motions were retrieved from the visual part of data. Style attributes like scene transitions, camera panning and zooming, music, and speech were derived from these properties. Finally, on the basis of style profile, video was classified as car racing, news, commercials, cartoon and tennis. Ramchandran et al. [14] proposed a consensus learning approach using YouTube categories for multi-label video categorization. Schindler et al. [15] categorized the videos using bag-of-words representation but the classification results are unsatisfactory. Wu et al. [16] used textual and social information for WVC that consists of user upload habits and YouTube related videos (specified by YouTube). Liu et al. [17] suggested a technique for video topic retrieval using 'related video' links that YouTube relates to each Web video to improve its textual information. Ballen et al. [18] proposed the use of social knowledge to suggest video tag and temporal localization. Besides using combination of modalities, many researchers depended just on text modality and used this information from social media, such as Wikipedia, and YouTube. Chen et al. [19] used Wikipedia categories (WikiCs) and content duplicated open resources (CDORs) for WVC.

2.1. Clustering ensemble

Clustering ensemble has become a decent approach while dealing with cluster analysis problems. Dimitriadou et al. [20] used the cluster alignment technique and proposed a voting based ensemble method. Fred et al. [21] considered the co-association matrix as the similarity matrix and proposed a new clustering ensemble technique. Strehl and Ghosh [22] presented an ensemble methodology based on hypergraph partitioning. Topchy et al. [23] extended the ensemble framework for generation of partitions by proposing an adaptive scheme for integration of multiple non-independent clustering. Fern and Brodley [24] solved the ensemble problem by converting it to a graph partitioning problem.

Recently, a semi-supervised clustering ensemble has been proposed and shown a remarkable performance by incorporating the known prior knowledge, e.g., pairwise constraints. Most commonly used constraints are ML and Cannot-Link (CL). Zhou and Li [25] proposed a semi-supervised learning paradigm based on disagreement, where multiple learners were trained for the task and the disagreements among the learners were exploited during the semi-supervised learning process. Zhou [26] explored that most semi-supervised ensemble methods proceeded by training learners using the initial training data first, and then using the learners to assign pseudo-labels to testing data. Iqbal et al. [27] used a voting scheme to solve semi-supervised clustering ensemble. Wang et al. [28] utilized the Bayesian network and EM algorithm to propose a semi-supervised cluster ensemble model. Yang et al. [29] exploited the multi-ant colonies to present a novel semi-supervised consensus clustering ensemble algorithm.

2.2. Semantic similarity

Semantic similarity is related to computing the similarity between concepts which are lexicographically dissimilar. There is a great deal of work in linguistics and computer science about using word (phrases) frequencies in text corpora to develop