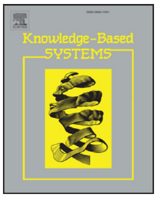
Knowledge-Based Systems 102 (2016) 64–75

Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Learning robust uniform features for cross-media social data by using cross autoencoders

Quan Guo[a], Jia Jia[b], Guangyao Shen[b], Lei Zhang[a,*], Lianhong Cai[b], Zhang Yi[a]

[a] College of Computer Science, Sichuan University, Chengdu, 610065, China
[b] Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Article history:
Received 19 November 2015
Revised 25 March 2016
Accepted 26 March 2016
Available online 29 March 2016

Keywords:
Cross-media
Social data
Cross modality
Deep learning
Autoencoder
Convolutional network

ABSTRACT

Cross-media analysis exploits social data with different modalities from multiple sources simultaneously and synergistically to discover knowledge and better understand the world. There are two levels of cross-media social data. One is the *element*, which is made up of text, images, voice, or any combinations of modalities. Elements from the same data source can have different modalities. The other level of cross-media social data is the new notion of *aggregative subject* (AS)— a collection of time-series social elements sharing the same semantics (*i.e.*, a collection of tweets, photos, blogs, and news of emergency events). While traditional feature learning methods focus on dealing with single modality data or data fused across multiple modalities, in this study, we systematically analyze the problem of feature learning for cross-media social data at the previously mentioned two levels. The general purpose is to obtain a robust and uniform representation from the social data in time-series and across different modalities. We propose a novel unsupervised method for cross-modality element-level feature learning called cross autoencoder (CAE). CAE can capture the cross-modality correlations in element samples. Furthermore, we extend it to the AS using the convolutional neural network (CNN), namely convolutional cross autoencoder (CCAE). We use CAEs as filters in the CCAE to handle cross-modality elements and the CNN framework to handle the time sequence and reduce the impact of outliers in AS. We finally apply the proposed method to classification tasks to evaluate the quality of the generated representations against several real-world social media datasets. In terms of accuracy, CAE gets 7.33% and 14.31% overall incremental rates on two element-level datasets. CCAE gets 11.2% and 60.5% overall incremental rates on two AS-level datasets. Experimental results show that the proposed CAE and CCAE work well with all tested classifiers and perform better than several other baseline feature learning methods.

1. Introduction

With the rapid development of the Internet, people have become increasingly dependent on social connections. Social media data are form by multiple modalities, for instance, text, images, voice, social interactions, *etc*. Moreover, the modalities in data samples vary very much. Social media data has created different types of correlational structures and distinctive statistical properties. Traditional approaches focus on dealing with single modality data or fusing data of multiple but same modalities. In contrast, cross-media learning focuses on homogeneous and heterogeneous multimedia data. This multimedia data from various sources needs to be integrated as a means to discover knowledge about the world synergistically. We refer to this problem as the core problem for cross-media learning.

Cross-media social data is based on two levels: the element level and aggregative level. At the element level, users create and spread numerous social media *elements* such as blogs, tweets, photos, and videos across various modalities. These blogs may have photos and videos often contain textual content such as hashtags and title descriptions; however, not every blog has a photo, and not every video includes text. At the aggregative level, collections of cross-media social elements are defined as *aggregative subjects* (AS) by the semantics they share. For example, photos make up an album on image-sharing websites such as Flickr or Instagram, where each album is an AS example; tweets make up a timeline for a user on social networks like Twitter or Facebook, where the timeline is an AS example; questions and comments make up a thread on Q&A communities like StackExchange or Quora where the thread is an AS example. Moreover, during an emergency event

[*] Corresponding author. Tel.: +862885400618; fax: +862885400618.
E-mail addresses: guoquanscu@gmail.com (Q. Guo), jjia@mail.tsinghua.edu.cn (J. Jia), thusgy2012@gmail.com (G. Shen), leizhang@scu.edu.cn (L. Zhang), clh-dcs@tsinghua.edu.cn (L. Cai), zhangyi@scu.edu.cn (Z. Yi).

http://dx.doi.org/10.1016/j.knosys.2016.03.028
0950-7051/© 2016 Elsevier B.V. All rights reserved.

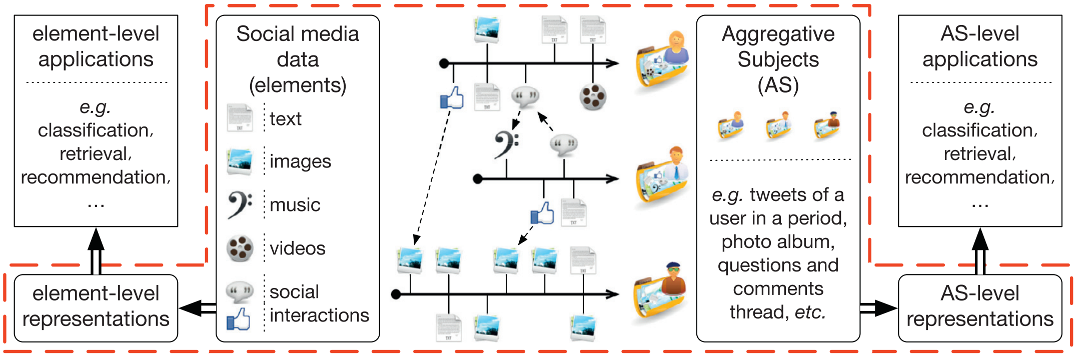

element-level applications
e.g. classification, retrieval, recommendation, …
element-level representations
Social media data (elements)
text
images
music
videos
social interactions
Aggregative Subjects (AS)
e.g. tweets of a user in a period, photo album, questions and comments thread, etc.
AS-level applications
e.g. classification, retrieval, recommendation, …
AS-level representations

Fig. 1. Illustration of the concepts and applications of learning robust and uniform features for cross-media social elements and AS.

like the Fukushima earthquake, a diverse set of people may upload photos, post tweets, and share blogs about this topic. For example, these social media data about the same Fukushima earthquake topic form an AS sample. Fig. 1 illustrates examples of social elements and AS. There are two characteristics of the elements in an AS: (1) they are in a time-series; (2) each element may contain multiple modalities. Nevertheless, the modalities of the elements may differ from each other. Elements have multiple and different modalities.

Given a social media dataset of elements or AS, the key problem to be addressed is establishing uniform features for unstructured homogeneous and heterogeneous cross-media data. There are certain demands for modeling cross-media social elements and AS for many social media applications including classification, retrieval, and recommendation systems. The representation of the data or the choice of features used to represent the inputs is critical to the overall performance of the applications.

In this study, our goal is to obtain robust features for cross-media social elements and simultaneously extract the uniform features for AS. The problem is non-trivial and poses a set of unique challenges. First, the elements are under a cross-modality setting. They can contain more than one modality. Moreover, their modalities can differ from each other. How do you obtain the modality-invariant representations? Second, the elements in AS are created over time and in time-series. Each of them has a specified context. How do you maximize the use of the time series and context information? Third, there are outliers among the elements of AS. Moreover, there are naturally occurring noise factors among the elements. For example, there can be document images such as “a passport” in a travel album. How to reduce the impact of outliers in data?

The red dashed-line box in Fig. 1 identifies the problem addressed in this study. The solid line with an arrowhead in the red box indicates the timeline of the elements in ASs. In addition, social elements are listed around the timeline. The dashed lines with the arrowheads indicate the targets of social interactions.

Deep learning [1,11,12], utilizing deep architectures and effective learning algorithms, has been emerging as a comprehensive paradigm for a vast range of problems. Krizhevsky et al. demonstrated a considerable improvement on image classification using convolutional neural networks (CNNs) [17]. Deep neural networks also achieve the state-of-the-art in multimedia areas with unstructured data [20,26,33]. Researchers also investigate neural networks for retrieval tasks [7,43]. There are works to integrate deep learning with other intelligence paradigms, for example, Zhou et al. [44] use deep neural networks for a context-aware stereotypical trust model in a multi-agent system.

We formulate the cross-media social elements feature learning problems and AS feature learning problems, respectively. We propose a novel unsupervised method for feature learning of cross-media social elements, namely cross autoencoder (CAE). A two-phase training method for training CAE with massive cross-modality data sample is presented. CAE can learn cross-modality correlation by an inductive cropping strategy, while also making use of the massive data with multiple and different modalities. Furthermore, we propose to use a CNN framework with CAE filters for AS-level feature learning, namely convolutional cross autoencoder (CCAE). We unroll the convolution operation and train CAE filters in CNN offline with the patches extracted from data samples. To the best of our knowledge, CCAE addresses a completely new problem to represent collections of cross-media elements, whereas previous technical works always focus on single independent elements [7,26,33].

Our contributions can be summarized as follows:

- We formulate the feature learning problem for cross-media social data with respect to social elements as well as social AS. We evaluate the quality of the learned features in the context of classification.
- We propose a CAE that learns modality-invariant features from cross-media social elements with different modalities in a two-phase unsupervised manner.
- Applying CAE as filters to handle cross-media elements, we employ a CNN framework to learn features for social AS. The CNN framework can manage the time sequence in social AS and reduce the impact of outliers in the social data.

To evaluate the quality of the proposed learning algorithm, we conduct experiments with classification tasks using real-world datasets from social media websites: Weibo, Sougo, and Flickr. We present experimental results for social elements using CAE and for social AS using CCAE. In terms of accuracy, CAE gets 7.33% and 14.31% overall incremental rates on two element-level datasets. CCAE gets 11.2% and 60.5% overall incremental rates on two AS-level datasets. Results indicate that CAE learns cross-modality correlation from cross-media social data. Further, supervised tasks using features from CAE show significant improvement as compared with baselines, and the experiments for AS show CCAE has superior performance for feature learning.

The remainder of this paper is organized as follows: In Section 2, we formulate the feature learning problem for cross-media social data. In Section 3, we briefly survey some mainstream methods for feature extraction and emphasize deep learning methods using autoencoders. In Section 4, we propose CAE for learning modality-invariant features of social media elements, and CCAE for learning uniform features for AS in social media. In Section 5, we present some experimental results. Section 6 concludes the paper.