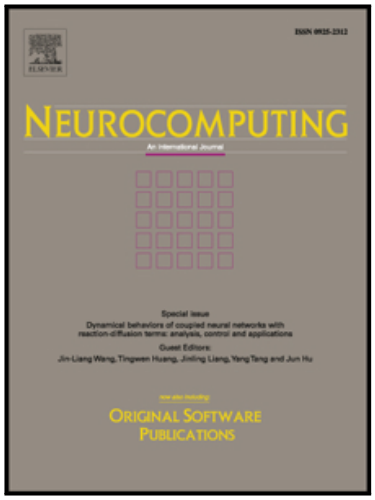
Communicated by Dr. Y. Chang

Accepted Manuscript



A Two-Step Approach to Describing Web Topics via Probable Keywords and Prototype Images from Background-removed Similarities

Junbiao Pang, Fei Tao, Liang Li, Qingming Huang, Baocai Yin, Qi Tian

PII: S0925-2312(17)31487-X
DOI: 10.1016/j.neucom.2017.08.057
Reference: NEUCOM 18858

To appear in: *Neurocomputing*

Received date: 26 February 2017
Revised date: 29 June 2017
Accepted date: 16 August 2017

Please cite this article as: Junbiao Pang, Fei Tao, Liang Li, Qingming Huang, Baocai Yin, Qi Tian, A Two-Step Approach to Describing Web Topics via Probable Keywords and Prototype Images from Background-removed Similarities, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.08.057

A Two-Step Approach to Describing Web Topics via Probable Keywords and Prototype Images from Background-removed Similarities

Junbiao Pang[a], Fei Tao[b], Liang Li[c], Qingming Huang[b,c], Baocai Yin[a,d], Qi Tian[e]

[a]*Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Faculty of Information Technology, Beijing University of Technology, No.100 Pingleyuan Road, Chaoyang District, Beijing 100124, China*
[b]*School of Computer and Control Engineering, University of Chinese Academy of Sciences, No.19 Yuquan Road, Shijingshan District, Beijing 100049, China*
[c]*Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, No.6 Kexueyuan South Road, Haidian District, Beijing 100190, China*
[d]*Dalian University of Technology, No.2 Linggong Road, Ganjingzi District, Dalian 116024, China*
[e]*Department of Computer Sciences, University of Texas at San Antonio, TX 78249, USA*

Abstract

To quickly grasp what interesting topics are happening on web, it is challenge to discover and describe topics from User-Generated Content (UGC) data. Describing topics by probable keywords and prototype images is an efficient human-machine interaction to help person quickly grasp a topic. However, except for the challenges from web topic detection, mining the multi-media description is a challenge task that the conventional approaches can barely handle: (1) noises from non-informative short texts or images due to less-constrained UGC; and (2) even for these informative images, the gaps between visual concepts and social ones. This paper addresses above challenges from the perspective of background similarity remove, and proposes a two-step approach to mining the multi-media description from noisy data. First, we utilize a devcovolution model to strip the similarities among non-informative words/images during web topic detection. Second, the background-removed similarities are reconstructed to identify the probable keywords and prototype images during topic description. By removing background similarities, we can generate coherent and informative multi-media description for a topic. Experiments show that the proposed method produces a high quality description on two public datasets.

Keywords: Topic Description, Poisson Deconvolution, User-Generated Content, Topic Detection, Background Similarity, Multi-modal Description

1. Introduction

With the rapid development of social media websites, the unprecedented explosion in the volume of User-Generated Content (UGC) data has made it difficult for web users to quickly grasp "hot" topics. Driven by such practical requirements, topic detection from web [1] [2] [3] is such an effort to organize web data into meaningful topics automatically. Formally, topic detection from web is defined as the task of discovering of a tiny fraction of interesting webpages strongly connected by a seminal event from a large amount of social media [1]. Even for the state-of-the-art methods [1] [4], a topic is typically detected as a cluster where a small number of webpages are uncorrelated to the theme of this topic. Naturally, the clustering-style representation barely supplies a "snapshot" way to help people to quickly understand the content of a topic. A naïve solution randomly samples a webpage as the prototype of a topic, facing the danger from these false detected webpages; besides, finding prototypes from noisy data is another open problem [4] [5]. Therefore, describing web topics from noisy multimedia data is a non-trivial task.

One of the important approaches is to generate multi-media description of a topic [6]. Generally speaking, the visual modality is very *vivid* but *indirect* for people to understand a topic [7]. As a contrast, the textual modality is more *semantic* but *unimpressive* than the visual one. Therefore, it is a natural way to represent a topic by the multi-modal representation: the visual modality supplies the vivid description, meanwhile the text one quickly pins down the semantic meaning of a topic. However, in the context of web topic detection, we argue that generating an *accurate* and *coherent* multi-media description should meet the two following challenges:

1) Non-informative words/images in UGC. These noises are essentially produced by the less-constrained "we meida", where data are posted at will across multiple modalities with few constraint. For instance, posted images are often uncorrelated to the content of a webpage;

2) Background words/images. Although these background words/images pave the way to exactly express the idea of a webpage, the background is not enough key to accurately describe a topic.

Email addresses: junbiao_pang@bjut.edu.cn (Junbiao Pang), fei.tao@vipl.ict.ac.cn (Fei Tao), liang.li@vipl.ict.ac.cn (Liang Li), qmhuang@ucas.ac.cn (Qingming Huang), ybc@dlut.edu.cn (Baocai Yin), qitian@cs.utsa.edu (Qi Tian)