Knowledge-Based Systems 108 (2016) 102–109

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Leveraging multimodal information for event summarization and concept-level sentiment analysis

Rajiv Ratn Shah[a,*], Yi Yu[b], Akshay Verma[c], Suhua Tang[d], Anwar Dilawar Shaikh[e], Roger Zimmermann[a]

[a] School of Computing, National University of Singapore, Singapore
[b] Digital Content and Media Sciences Research, National Institute of Informatics, Japan
[c] Department of Computer Science and Engineering, MNNIT, India
[d] Graduate School of Informatics and Engineering, UEC, Japan
[e] Department of Computer Engineering, Delhi Technological University, India

ARTICLE INFO

Article history:
Received 15 November 2015
Revised 7 May 2016
Accepted 10 May 2016
Available online 11 May 2016

MSC:
00-01
99-00,

Keywords:
Multimedia summarization
Semantics analysis
Sentics analysis
Multimodal analysis
Multimedia-related services

ABSTRACT

The rapid growth in the amount of user-generated content (UGCs) online necessitates for social media companies to automatically extract knowledge structures (concepts) from photos and videos to provide diverse multimedia-related services. However, real-world photos and videos are complex and noisy, and extracting semantics and sentics from the multimedia content alone is a very difficult task because suitable concepts may be exhibited in different representations. Hence, it is desirable to analyze UGCs from multiple modalities for a better understanding. To this end, we first present the EventBuilder system that deals with semantics understanding and automatically generates a multimedia summary for a given event in real-time by leveraging different social media such as Wikipedia and Flickr. Subsequently, we present the EventSensor system that aims to address sentics understanding and produces a multimedia summary for a given mood. It extracts concepts and mood tags from visual content and textual metadata of UGCs, and exploits them in supporting several significant multimedia-related services such as a musical multimedia summary. Moreover, EventSensor supports sentics-based event summarization by leveraging EventBuilder as its semantics engine component. Experimental results confirm that both EventBuilder and EventSensor outperform their baselines and efficiently summarize knowledge structures on the YFCC100M dataset.

1. Introduction

The number of UGCs (*e.g.*, photos and videos) has increased dramatically in recent years due to the ubiquitous availability of smartphones, digital cameras, and affordable network infrastructures. An interesting recent trend is that social media companies such as Flickr and YouTube, instead of producing content by themselves, create opportunities for a user to generate multimedia content. Thus, capturing the multimedia content anytime and anywhere, and then instantly sharing them on social media platforms, has become a very popular activity. Since UGCs belong to different interesting events (*e.g.*, festivals, games, and protests), they are now an intrinsic part of humans' daily life. For instance, on the very popular photo sharing website Instagram, over one billion photos have been uploaded so far. Moreover, the site has more than 400 million monthly active users [1]. However, it is difficult to automatically extract knowledge structures from the multimedia content due to the following reasons: (i) the difficulty in capturing the semantics and sentics of UGCs, (ii) the existence of noise in textual metadata, and (iii) challenges in handling big datasets. First, aiming at the understanding of semantics and summarizing knowledge structures of multimedia content, we presented the EventBuilder[1] system in our earlier work [2]. It enables users to automatically obtain a multimedia summary for a given event from a large multimedia collection in real-time (see Fig. 1). This system leverages information from social media platforms such as Wikipedia and Flickr to provide a useful summary of the event. Since this earlier work mainly focused on a real-time demonstration, its performance evaluation was limited. Thus, in this study, we perform extensive experiments of EventBuilder on a collection of 100 million photos and videos from Flickr and compare the

* Corresponding author.
E-mail address: rajiv@comp.nus.edu.sg (R.R. Shah).

[1] http://www.yiyu.nii.ac.jp:8080/EventBuilder/

http://dx.doi.org/10.1016/j.knosys.2016.05.022
0950-7051/

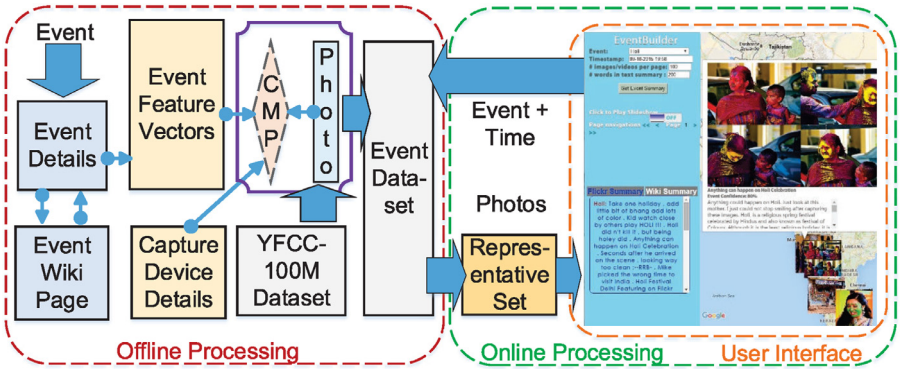

Fig. 1. System framework of the EventBuilder system.

results with a baseline. In the baseline system, we select photos that contain the input event name in their metadata (*e.g.*, descriptions, titles, and tags). Experimental results confirm that the proposed algorithm efficiently summarizes knowledge structures and outperforms the baseline. Next we describe how our approach solves above mentioned problems.

Advancements in technologies have enabled mobile devices to collect significant amounts of contextual information (*e.g.*, spatial, temporal, and other sensory data) in conjunction with UGCs. We argue that the multimodal analysis of UGCs is very helpful in semantics and sentics understanding because often the multimedia content is unstructured and difficult to access in a meaningful way from only one modality. Since multimodal information augments knowledge bases by inferring semantics from the unstructured multimedia content and contextual information, we leverage it in the EventBuilder system, which has the following three novel characteristics: (i) leveraging Wikipedia as event background knowledge to obtain additional contextual information about an input event, (ii) visualizing an interesting event in real-time with a diverse set of social media activities, and (iii) producing text summaries for the event from the description of photos and Wikipedia texts by solving an optimization problem.

Next, aiming at understanding sentiments and producing a sentics-based multimedia summary from a multimedia collection, we introduce the EventSensor[2] system in this study. Moreover, EventSensor leverages EventBuilder as its semantics engine to produce sentics-based event summarization. EventSensor leverages multimodal information for sentiment analysis from UGCs. It extracts concepts from the visual content and textual metadata of a photo and exploits them to determine the sentics details of the photo. A concept is a knowledge structure which provides important cues about sentiments. For instance the concept *"grow movement"* indicates anger and struggle. Textual concepts (*e.g.*, grow movement, fight as community, and high court injunction) are computed from the textual metadata such as description and tags by the semantic parser provided by Poria *et al.*[3]. Visual concepts are tags derived from the visual content of photos and videos by using a convolutional network that indicate the presence of concepts such as people, buildings, food, and cars. The visual concepts of all photos in the YFCC100M dataset are provided as metadata. On this bases, we propose a novel algorithm to fuse concepts derived from the textual and visual content of a photo (see Algorithm 1). Subsequently, we exploit existing knowledge bases such as SenticNet-3, EmoSenticNet, EmoSenticSpace, and WordNet to determine to the sentics details of the photo. Such knowledge bases help us to build a sentics engine which is helpful in providing sentics-based services. For instance, the sentics engine is

Algorithm 1 Fusion of concepts.

1: **procedure** ConceptFusion
2: INPUT: Textual concepts C_T and visual concepts C_V of a photo p
3: OUTPUT: A list of fused concepts C for p
4: $C = \emptyset$ ▷ initialize the set of fused concepts for p.
5: **if** (hasTags(p) == *TRUE*) **then** ▷ check if p has tags.
6: C = getTagConcepts(C_T) ▷ Since tags has the highest accuracy, see Fig. 6.
7: **else if** ((hasDescription(p) ∧ hasVisualConcepts(p)) == *TRUE*) **then**
8: C = getDescriptionConcepts(C_T) ▷ get concepts from descriptions.
9: $C = C \bigcup C_V$ ▷ Since it has the second highest accuracy, see Fig. 6.
10: **else if** ((hasTitle(p) ∧ hasVisualConcepts(p)) == *TRUE*) **then**
11: C = getTitleConcepts(C_T) ▷ get concepts from descriptions.
12: $C = C \bigcup C_V$ ▷ Since it has the third highest accuracy, see Fig. 6.
13: **else if** (hasVisualConcepts(p) == *TRUE*) **then**
14: $C = C_V$ ▷ Since it has the fourth highest accuracy, see Fig. 6.
15: **else if** (hasDescription(p) == *TRUE*) **then** ▷ check if p has description.
16: C = getDescriptionConcepts(C_T) ▷ Since it has 5th highest accuracy.
17: **else if** (hasTitle(p) == *TRUE*) **then** ▷ check if p has title.
18: C = getTitleConcepts(C_T) ▷ Since it has the lowest accuracy.
19: return C ▷ A set of fused concepts for p.

used for a mood-related soundtrack generation in our system (see Fig. 2). Mood-based sound that matches with emotions in multimedia content is a very important aspect and contributes greatly to the appeal of a video when it is being viewed. Thus, a video (*i.e.*, a slideshow of photos) with a matching soundtrack has more appeal for viewing and sharing on social media websites than a normal slideshow of photos without an interesting sound. This inspires people to often create such a music video by adding matching soundtracks to photos and share them on social media. However, adding soundtracks to photos is not easy due to the following reasons. Firstly, traditionally it is tedious, time-consuming, and not scalable for a user to add custom soundtracks to photos from a large collection of UGCs. Secondly, it is difficult to extract moods of photos automatically. Finally, an important aspect is that a good soundtrack should match and enhance the overall moods

[2] http://pilatus.d1.comp.nus.edu.sg:8080/EventSensor/