World Wide Web (2016) 19:177–179
DOI 10.1007/s11280-015-0360-2

Guest editorial: web multimedia semantic inference using multi-cues

Yahong Han[1,4] · **Yi Yang**[2] · **Xiaofang Zhou**[3]

Published online: 23 June 2015

With the popularity of social media applications and Web 2.0 techniques, user-generated multimedia contents such as blogs, test messages, photos, videos, user click log and Place of Interest (POI) check-ins become pervasive, which enables the study on exploiting them as multiple cues for web multimedia semantic inference. Most of the time when one speaks of web multimedia corpora, he/she may think of heterogeneous corpora consisting of data from various sources, and of different modality. The heterogeneous multimedia content provides a variety of cues for semantic inference of real-world multimedia applications. Research so far has mostly focused on mono-cue analysis of multimedia content, such as looking only into images, videos, or text, but rarely leverage multiple semantic cues like the surrounding texts of images/videos on a web page or the click logs of users' profiles from the same community. As such, new algorithms and models for analyzing correlations among multiple semantic cues become one of the most active research areas in web multimedia applications.

From the above background, many efforts have focused on the utilization of multiple semantic cues for web multimedia semantic inference. Particularly, the different semantic cues may be temporally synchronized (e.g., video clips and corresponding audio transcripts, animations, multimedia presentations), spatially related (images embedded in text, object relationships in 3D space), semantically correlated (combined analysis of collections of videos, set of images created by one's social network), or otherwise click-through connected (images

✉ Yahong Han
yahong@tju.edu.cn

Yi Yang
yee.i.yang@gmail.com

Xiaofang Zhou
zxf@itee.uq.edu.au

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China

[2] Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia

[3] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

[4] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China

with search engine's user click log). The correlations provide abundant context information for web multimedia semantic analysis.

This special issue has gained overwhelming attention and received 23 submissions from researchers and practitioners working on Web multimedia semantic analysis. After initial examining of all submissions, 21 papers are selected into the regular rigorous review process and each submission has been reviewed by at least two reviewers. After 2–3 round reviews, eventually seven quality papers are recommended to be included into this special issue, which are summarized as below.

Multi-modality and cross-media analysis are typical ways of multi-cue analysis in Web multimedia semantic inference. In this issue, two papers investigate cross-media distance metric learning and domain adaptation for Web multimedia semantic analysis. The paper, titled "A Cross-media Distance Metric Learning Framework based on Multi-view Correlation Mining and Matching", presents a novel cross-media distance metric learning framework based on sparse feature selection and multi-view matching. Cross-media distance metric learning focuses on correlation measure between multimedia data of different modalities. However, the existence of content heterogeneity and semantic gap makes it very challenging to measure cross-media distance. The proposed method employs sparse feature selection to select relevant features and then maximize the canonical coefficient during image-audio feature dimension reduction for cross-media correlation mining. A multi-modal semantic graph is constructed to find the embedded manifold cross-media correlations. The proposed method shows good performance in cross-media retrieval for image-audio dataset. Different from the cross-media analysis, domain adaptation aims at exploring how to improve the semantic analysis of target domain by transferring knowledge from related auxiliary domains. The paper "Active Domain Adaptation with Noisy Labels for Multimedia Analysis" targets the fact that automatic the Web semantic detector may be noisy and the number of positive examples is limited. For example, available labeled data are insufficient in many applications. Since the source and target domains usually have different distributions, it is possible that the domain expert may not have sufficient knowledge to answer each query correctly. The paper proposes to utilize active learning and domain adaptation to minimize the required amount of labeled data for model training. The proposed method can efficiently minimize the required amount of labeled data for domain adaptation. Experiments on the tasks of Web image classification and cross-domain head pose estimation in videos show the effectiveness of the proposed method.

Feature construction or feature learning is a fundamental problem in Web multimedia semantic analysis, as it can help bridge the semantic gap between low-level features and high-level semantic. The paper "Feature Aggregating Hashing for Image Copy Detection" and "Compact Representation for Large-Scale Unconstrained Video Analysis" explore how to effectively represent images or videos for image copy detection and video recognition. The paper "Feature Aggregating Hashing for Image Copy Detection" and "Compact Representation for Large-Scale Unconstrained Video Analysis" proposes a promising approach using binary finger-prints to define visual words. Then machine learning based hashing to generate binary codes of visual words and get hashing functions which map local features into binary code efficiently. Histograms of visual words are constructed as image representation and histogram comparison is employed to measure the similarity between two images. Experiments on benchmark image datasets demonstrate the better performance of the proposed method in image copy detection. As most of the recent proposed video features are in high dimensionality, the computation costs are increased especially in the large-scale Web applications. The paper "Compact Representation for Large-Scale Unconstrained Video Analysis"

targets to construct effective and compact representation for large-scale video recognition. The paper proposes a novel semi-supervised feature selection algorithm to reduce redundant feature dimensions. Different from most of the existing semi-supervised feature selection algorithms, the proposed algorithm does not rely on manifold approximation, which is quite expensive for a large number of data. Thus, it is possible to apply the proposed algorithm to a real large-scale video analysis system.

In Web multimedia retrieval, users' search intention is an important semantic cue, which is always referred to in the so-called semantic gap between low level features and search intention. Different from the key-word search, complex queries may convey more specific searching needs. The paper "Complex-Query Web Image Search with Concept-Based Relevance Estimation" proposes a new image re-ranking scheme based on concept relevance estimation. The proposed method utilizes probabilistic models to account for the correlations between concept and query as well as between concept and images. In each model, different cues like visual, web, and textual information are incorporated. Moreover, the scheme also takes multiple Web sources as sampling source of labeled images. Experiments on complex-query Web image retrieval show the better performance of the proposed method. Results from real-word dataset collected from WikiAnswers demonstrate the proposed method has the potential of real-world applications.

Benchmark datasets construction and the development of evaluation methods are important to the research of Web multimedia semantic inference. Most of the benchmark datasets are constrained, e.g., for videos, with few variations of scenes, viewpoints, background clutter, which is too simple to promote more robust approaches. The paper "From constrained to unconstrained datasets an evaluation of local action descriptors and fusion strategies for interaction recognition" introduce a new unconstrained video dataset for interaction recognition. Compared to existing datasets, the constructed dataset represents realistic scenes and has more challenges of different variations of scenes, viewpoints, background clutter, imaging platforms etc. Besides, multiple widely-employed low-level feature descriptors are extracted. They also conduct evaluation studies and comparison experiments on the dataset. The results show the potential of the dataset to promote practical methods on interaction video recognition.

We also include in this issue a real-world application of social multimedia mining and knowledge discovering using multi-cues. The paper "Sub-Event Discovery and Retrieval during Natural Hazards on Social Media Data" proposes a new method for sub-events discovery in natural hazard, which adopts multifarious features to detect sub-events. In order to retrieve the sub-events over a specific event, they introduce a novel SER (Sub-Event Retrieval) algorithm from time-stamped social media data. The proposed SER makes use of automatically obtained messages from external search engines in the entire process. In the experiments, they collected real-world data from a social multimedia micro-blog website Sina Weibo. The comparison results show the better performance of the algorithm. Besides, they gave an example and detailed analysis of conducting the proposed method for a particular "Earthquake" event on the Sina Weibo website.

Finally, we would like to appreciate all authors who submitted manuscripts for consideration, and anonymous dedicated reviewers for their criticism and time to help us making final decisions. Without their valuable and strong supports, we cannot make this special issue successful. Our sincere gratitude will also go to the WWWJ EiC, Prof. Yanchun Zhang, Ms. Jennylyn Roseiento, and Ms. Melissa Fearon from the Springer Journal Editorial Office for helping us to presenting this special issue to readers.