Multi-Modal Learning over User-Contributed Content from Cross-Domain Social Media

Wen-Yu Lee
National Taiwan University, Taipei, Taiwan
majorrei@cmlab.csie.ntu.edu.tw

Abstract

The goal of the research is to discover and summarize data from the emerging social media into information of interests. Specifically, leveraging user-contributed data from cross-domain social media, the idea is to perform multi-modal learning for a given photo, aiming to present people's description or comments, geographical information, and events of interest, closely related to the photo. These information then can be used for various purposes, such as being a real-time guide for the tourists to improve the quality of tourism. As a result, this research investigates modern challenges of image annotation, image retrieval, and cross-media mining, followed by presenting promising ways to conquer the challenges.

Every day, millions of media data are uploaded to social sharing websites. It is desirable to discover what the large number of data can be used to make our lives better. In our daily life, "what is this?" and "where am I?" are two common questions that arise when people go out. Earlier, it could be troublesome or even hard to find the answers, especially when there is a language barrier, e.g., when travelling abroad. Moreover, people are often like to explore more about the places they are visiting. "Are there special events or festivals?" "What happened?" It is interesting to discover a specific region from various perspectives, such as foods and clothing.

As a solution, my dissertation attempts to build an effective social-media mining system for these questions, by developing effective techniques of image annotation, image retrieval, and cross-media mining, considering modern challenges. Figure 1 illustrates the idea. Ideally, people may upload a photo onto the system. The system then performs social-media information retrieval based on the photo. Specifically, the system is expected to generate textual tags that reflect the image content of the photo by image annotation, geo-tags that indicates where the photo might be taken by image retrieval, and events of interests and/or various perspectives related to the photo by cross-media mining.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

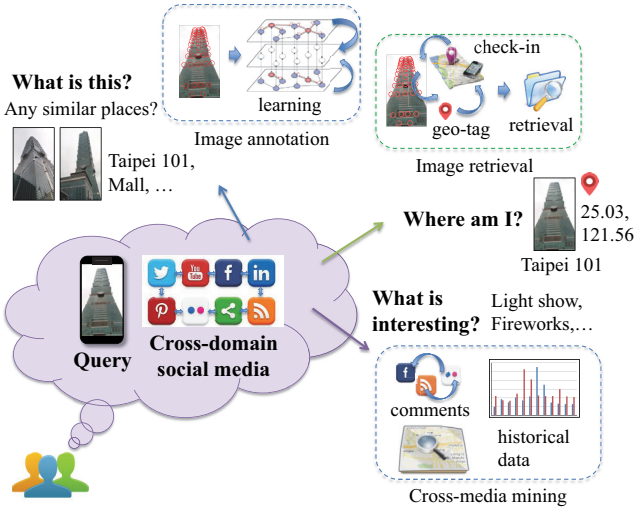

Figure 1: Illustration of the desired social-media mining system, including the components of image annotation, image retrieval, and cross-media mining.

Problem Formulation

Although image annotation, image retrieval, and media mining, are classical problems, modern challenges have reshaped the problems. We focus on the following challenges of these problems.

- Image annotation: graph-based semi-supervised learning is a powerful technique to *propagate* tags from tagged images to un-tagged images, where handling large-scale image datasets and high-dimensional data are critical for modern tag propagation.
- Image retrieval: most previous works were based on image clustering of visual features and geo-tags; however, geo-tags of images might not be accurate enough and may be limited on distinguishing buildings in a close proximity. In addition, images of image clusters may be noisy.
- Cross-media mining: most previous works summarized events a single media stream, where the information could be limited because different media streams have different features. It is desirable to combine multiple media streams

for high performance and different perspectives.

Multi-tag propagation on large-scale datasets

Extended a prior work on single-tag propagation (Zhou et al. 2004), I developed a scalable multi-tag propagation approach using the MapReduce programming model, so as to confront large-scale datasets with high-dimensional data (Lee et al. 2013). Further, unlike traditional ad-hoc approaches, I determined the number of iterations for iterative learning by formulating an all-pair shortest path problem for a better performance. Finally, I extended the weight learning technique (Jiang, Wang, and Chang 2011) for the refinement of tag assignment, considering multiple similarity graphs of images. An experiments was conducted on a large-scale image dataset that contains more than 540,000 images with 9,360 annotated ground truth images in 21 query categories. Results showed that my approach improved about 77% mean average precision for top-10 tags, compared with a common approach considering only visual features.

Further, we discovered relevance of tags and images (Wu et al. 2013).[1] Future work includes the improvement of the approach and then uses the approach to identify noisy tags generated during the learning process for adaptive learning on tag propagation.

City-view image retrieval using check-in data

Our preliminary study presented a real-time retrieval system leveraging visual features and geo-tags (Kuo et al. 2011).[2]

To remedy the deficiency of geo-tags, to the best of our knowledge, I further presented the *first work* that unifies visual features, geo-tags, and check-in data of images for city-view image retrieval (Lee, Kuo, and Hsu 2013). Check-in data were regarded as complementary location information of images, in contrast to geo-tags. For data pre-processing, check-in data were used for initial image clustering. Sparse coding with effective dictionary selection was then used to discover critical feature dimensions of images for each cluster. For image retrieval, a ranked list of images will be generated based on a given photo. The check-in location and geo-tags of clusters can be used to indicate where the given photo may be taken. Experiments on an image dataset of San Francisco showed that the approach outperformed common approaches that considers visual features only, geo-tags only, and both of visual features and geo-tags.

Recently, I observed that some check-in data may be noisy. Some photos might not be closely related to the places these photos were taken at, but related to other places. For example, image contents of photos taken at a skyscraper may be more closely related to buildings near the skyscraper than the skyscraper itself. As a result, I have developed a location-aware regrouping approach to optimize initial clusters for search performance improvement. By the workshop date, the regrouping approach will be implemented and evaluated based on two or more image datasets.

Cross-media mining for events of interest

Previous works, such as (Sakaki, Okazaki, and Matsuo 2010), mostly focused on mining events of interest from a single media stream. Intuitively, unifying multi media streams is capable of achieving better performance than considering one stream alone, because every media stream has unique information and features. As a case study, I conducted an experiment in our preliminary work (Kuo et al. 2014)[3] that compared top-100 popular places from different media streams, including Twitter, Instagram, TripAdvisor, and one NYC open data. As expected, each media stream has its own unique information and features, and it is desirable to do cross-domain combination.

In addition, I led a team to investigate what people may concern with for different types of events, by extracting latent sub-events with diverse and representative attributes for pre-selected events (Lee et al. 2015).[4] Currently, I am working on unifying information of social websites and that of taxi datasets to detect events of interest in a city. By the workshop date, I will evaluate the performance of my current approach and refine the approach accordingly.

References

Jiang, Y.-G.; Wang, J.; and Chang, S.-F. 2011. Lost in binarization: query-adaptive ranking for similar image search with compact codes. In *ACM ICMR*.

Kuo, Y.-H.; Lee, W.-Y.; Hsu, W.; and Cheng, W.-H. 2011. Augmenting mobile city-view image retrieval with context-rich user-contributed photos. In *ACM MM*, 687–690.

Kuo, Y.-H.; Chen, Y.-Y.; Chen, B.-C.; Lee, W.-Y.; Wu, C.-C.; Lin, C.-H.; Hou, Y.-L.; Cheng, W.-F.; Tsai, Y.-C.; Hung, C.-Y.; Hsieh, L.-C.; and Hsu, W. 2014. Discovering the city by mining diverse and multimodal data streams. In *ACM MM*, 201–204.

Lee, W.-Y.; Hsieh, L.-C.; Wu, G.-L.; and Hsu, W. 2013. Graph-based semi-supervised learning with multi-modality propagation for large-scale image datasets. *Elsevier JVCIR* 24(3):295–302.

Lee, W.-Y.; Kuo, Y.-H.; Hsieh, P.-J.; Cheng, W.-F.; Chao, T.-H.; Hsieh, H.-L.; Tsai, C.-E.; Chang, H.-C.; Lan, J.-S.; and Hsu, W. 2015. Unsupervised latent aspect discovery for diverse event summarization. In *ACM MM*, 197–200.

Lee, W.-Y.; Kuo, Y.-H.; and Hsu, W. 2013. City-view image retrieval leveraging check-in data. In *ACM GeoMM*, 13–18.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *ACM WWW*, 321–328.

Wu, C.-C.; Chu, K.-Y.; Kuo, Y.-H.; Chen, Y.-Y.; Lee, W.-Y.; and Hsu, W. 2013. Search-based relevance association with auxiliary contextual cues. In *ACM MM*, 393–396.

Zhou, D.; Bousquet, Q.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2004. Learning with local and global consistency. In *NIPS*, 321–328.

[1]This work received the first place of ACM Multimedia 2013 MSR-Bing Image Retrieval Challenge (Industrial Track), and the ACM Multimedia 2013 Grand Challenge Multimodal Award.

[2]This work was accepted as the finalists for ACM Multimedia 2011 Grand Challenge.

[3]This work received the ACM Multimedia 2014 Grand Challenge Multimodal Award.

[4]This work was accepted as the finalists for ACM Multimedia 2015 Grand Challenge.