Knowledge Discovery from Community-Contributed Multimedia

Tao Mei
Microsoft Research Asia

Winston H. Hsu
National Taiwan University

Jiebo Luo
Kodak Research Laboratories

The prevalence of image- and video-capturing devices and the advent of media-sharing services such as Flickr and YouTube have drastically increased the volume of community-contributed multimedia. For example, there are reportedly more than four billion images in Flickr and 24 hours of new videos are uploaded to YouTube every minute. Such a vast amount of photos, videos, and music shared via websites is bound to exert a profound social impact on human society and poses a new challenge for developing efficient indexing, search, mining, and visualization approaches for managing such large-scale media data.

Social media is augmented with rich context, such as user-provided tags, comments, geolocations, time, device metadata, and so on. These bits of context data are promising resources to exploit for benefiting a wide variety of applications, such as annotation, recommendation, questioning and answering, advertising, and (cultural) activity discovery.

The goal of this special issue is to present a concise reference of state-of-the-art efforts in such attempts for knowledge discovery over large-scale, community-contributed multimedia, and in particular the opportunities and challenges in this nascent arena. We have selected five articles that represent ways to exploit user-contributed photos and videos for several applications and that identify the theoretical challenges associated with managing such multimedia data.

The articles

Among the many forms of rich context associated with media, tags are used most in a wide variety of applications (for example, search, organization, recommendation, and visualization). However, large-scale, weakly tagged images might suffer seriously from the problem of tag uncertainty, which in turn can prevent users from being able to use the application effectively. To begin with, Fan et al. briefly survey the current attempts in improving tagging quality associated with images and videos. They propose a cross-modal, tag-cleansing algorithm that integrates the visual similarity contexts of the weakly tagged images with the semantic similarity contexts of their tags.

Rather than watching the broadcasts alone, sports fans are keen to share their passions through social media (for example, Facebook, YouTube, blogs, and Twitter). Traditionally in the multimedia research community, sports event detection (or summarization) is performed through supervised or unsupervised content analysis. The next article, by Smits and Hanjalic, augments common content-analysis approaches with collaborative (community) tagging. The two aspects compensate for the shortcomings of each other: automatic content analysis helps guide users toward potentially interesting highlight candidates, while tagging itself validates and enriches the detected results through a collaborative effort. This naturally brings the motivated fans into the loop.

The rich context of social media, such as user-provided tags and geolocations, creates rich sets of georeferenced photos. Newsam presents a research survey in the third article that focuses on the knowledge discovery in effective crowdsourcing what-is-where on the surface of the earth. In particular, the author illustrates recent results of leveraging large collections of georeferenced photos to solve three

1070-986X/10/$26.00 © 2010 IEEE

Published by the IEEE Computer Society

classes of problems: annotating novel images, annotating geographic locations, and performing geographic discovery. He also argues that such contextually rich collections can be used as an alternative for investigating physical, cultural, and behavioral aspects of the earth.

In recent years, we have witnessed the prevalence of community-based question-answering (QA) systems that are able to provide precise answers to a wide variety of questions. However, answers from most QA systems are in the form of text. For some questions, especially with how-to questions, video answers would be more effective and intuitive, particularly in light of such questions as how to transfer pictures in a digital camera to a computer. Li et al. present a solution to the how-to QA by leveraging the community-contributed comments and video answers on the Web. They demonstrate the feasibility by using YouTube videos to solve several questions about consumer electronics.

The last article, by Wu, Ngo, and Zhao, argues that, along with the sheer amount of multimedia data, conventional model-based approaches normally involve a large set of classifiers and become impractical due to the scarcity of (manually collected) training examples, complexity in learning, poor generalization across domains, and so on. By describing how to leverage freely available user-contributed data in a data-driven manner, they show that quite a few problems (that is, near-duplicate detection, Web video annotation, and video categorization) can be solved without the need for sophisticated algorithms when exploiting rich contextual and social resources such as filters, features, training data, and domain knowledge. They further demonstrate that such data-driven approaches are easy to deploy and promisingly scalable.

Further challenges and directions

Community-contributed photos, videos, and comments have proved a valuable resource for discovering knowledge and further developing promising applications. In the initial research stage, there are several open issues worth further study in this area:

- effective and efficient content-based image and video retrieval in billion-scale collections, especially in the situation where the objects of interest (for example, people, products, landmarks, and so on) only occupy a small portion of the images or the videos;
- scaling the statistical learning algorithms to deal with Web-scale training data (for example, leveraging the distributed computation platform);
- managing and mitigating the inaccuracy and incompleteness of user-contributed annotation for data-driven methods; and
- analyzing and transferring training data from one specific domain to others (for example, from YouTube videos to Flickr photos, or vice versa).

Besides the sample applications mentioned in this special issue, there are many other promising ones, including visualization on unstructured images and videos, context-aware multimedia advertising, and augmented reality for bridging the gap between social interaction and the physical world. **MM**

Tao Mei is a research staff member at Microsoft Research Asia. His research interests include multimedia content analysis; computer vision; and Internet media applications, such as search, advertising, recommendation, presentation, and social network. Mei has a PhD in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China. Contact him at tmei@microsoft.com.

Winston H. Hsu is an assistant professor in the Graduate Institute of Networking and Multimedia and the Department of Computer Science and Information Engineering, National Taiwan University, and the founder of the Multimedia Indexing, Retrieval, and Analysis Research Group. His research interests include semantic understanding, mining, and retrieval over large-scale image and video collections. Hsu has a PhD in electrical engineering from Columbia University, New York. Contact him at winston@csie.ntu.edu.tw.

Jiebo Luo is a senior principal scientist with the Kodak Research Laboratories, Rochester, New York. His research interests include signal and image processing, pattern recognition, computer vision, multimedia data mining, biomedical informatics, computational photography, human–computer interaction, and ubiquitous computing. Luo has a PhD in electrical engineering from the University of Rochester. He has authored over 160 papers and 60 US patents, and is a Fellow of IEEE, SPIE, and IAPR. Contact him at jiebo.luo@kodak.com.