Guest Editors' Introduction

Large-Scale Multimedia Retrieval and Mining

Rong Yan
Facebook

Benoit Huet
EURECOM

Rahul Sukthankar
Intel Labs and Carnegie Mellon University

Recent years have witnessed an explosive growth of multimedia data due to higher processor speeds, faster networks, wider availability of high-capacity mass-storage devices, and the advent of cloud computing. Stimulated by current work in scalable machine learning, feature indexing and multimodal analysis techniques, researchers are increasingly interested in exploring challenges and new opportunities for developing scalable approaches for multimedia retrieval and mining. The enormous scale of multimedia data is reflected in the following statistics: approximately 120 million digital still cameras were sold in 2010; video already accounts for more than half of all Internet traffic, with YouTube attracting more than 2 billion views per day and 24 hours of video uploaded every minute. This explosion of the amount of data, number of users, and availability of new resources has led to greater expectations for multimedia retrieval and mining in terms of effectiveness and efficiency, for which existing analysis approaches and systems typically don't suffice.

Scalability to large data collections poses a particularly significant challenge for current multimedia processing methods. For instance, only about one-third of the papers appearing in ACM Multimedia 2008's content track are applicable to this scale of data collections. Meanwhile, the research interest in processing large-scale image, audio, and video collections continues to grow rapidly, given the increasing availability of Web 2.0 websites, surveillance videos, and both personal and enterprise multimedia archives. We believe the tipping point for large-scale multimedia analysis is quickly approaching.

This special issue samples the state of the art in large-scale multimedia analysis techniques and explores how advanced multimedia analysis can be leveraged to address the challenges in large-scale data collections. In particular, from a total of 20 submissions, we selected five representative articles, that investigate large-scale multimedia analysis theory and systems across multiple application domains, such as Web event detection, landmark detection, image annotation, musical content mining, and cloud computing.

Summary of articles

Shavitt, Weinsberg, and Weinsberg's article describes a large-scale study of song files shared by users in the Gnutella network. The authors observe that mining peer-to-peer (P2P) networks is challenging due to the high degree of user churn and significant noise in the user-generated metadata. To address this challenge, the article presents a method for automatically constructing song similarity that allows them to assess the distance between and like-mindedness of users. To mitigate the effects of noise and missing metadata, the authors propose graph *k*-medoids (GkM), a scalable variant of *k*-medoid clustering designed to identify groups of similar songs more effectively than popular unsupervised clustering algorithms. GkM scales to large-scale graphs due to specific adaptations, such as employing a single iteration of medoid selection, thus sacrificing optimality for scalability, and operating directly on the similarity graph rather than requiring the data to be projected into a metric space. While this study focuses on musical content in Gnutella, the approach described in the article is broadly applicable to noisily tagged, large-scale media collections.

A standard approach to large-scale content search of unannotated images is to represent

1070-986X/11/$26.00 © 2011 IEEE Published by the IEEE Computer Society

semantic information using a bag-of-words (BoW) model. Unfortunately, the typical BoW codebook-generation process results in a significant loss of sematic information. To overcome this drawback, Wu and Hoi propose an online semantics-preserving, metric-learning algorithm for enhancing BoW by minimizing the semantic loss. Derived from the batch algorithm on semantics-preserving metric learning, this online counterpart is designed to be more efficient and scalable for large-scale, image-annotation applications. The authors demonstrate the effectiveness of the proposed methods on two large-scale data collections including Flickr and LabelMe.

With a rapidly growing volume of online video, it becomes increasingly important to summarize the event highlights with keywords and images so users can effectively browse through video retrieval results. To illustrate the event evolution for video, Wu et al. explore the issues of event discovery and structure construction by analyzing text co-occurrence and visual near-duplicate feature trajectory. On the basis of event similarity, they built an event structure by linking and aligning events along a timeline. They associated representative text keywords and visual frames to each event to enable effective visualization and browsing of Web videos.

Events and landmarks are important criteria for organizing photo collections. In their work, Papadopoulos et al. propose a detection scheme for annotating images using hybrid graph-based clustering that operates on both the image annotation and its visual content. The clusters are obtained using community detection, where graph nodes with high connectivity are grouped together. The approach, evaluated on a large collection originating from Flickr, shows that both landmarks and events can be extracted with high precision.

Scalability is a major issue when dealing with the massive quantities of multimedia information produced and uploaded on the internet. Candan et al. propose a parallel processing system, RanKloud, which prunes out unpromising media items so that no or limited computational resources are wasted. The system, based on MapReduce, is designed to allow great scalability when used on clouds. It extends the key/value pair with a utility score that indicates the relevance and importance of the data. The study shows that performing runtime statistics on the data allows for improving large-scale operations by avoiding waste processing and balanced partitioning.

Future directions

We envision several future opportunities in the area of large-scale retrieval and mining that are worthy of attention from the multimedia community:

Many technical issues are yet to be addressed when managing large multimedia collections, for example, how to obtain accurate annotations, how the visual features can be efficiently indexed, how best to create large-scale benchmark collections, and how to organize these annotations.

Most state-of-the-art, machine-learning algorithms, such as nonlinear kernel support vector machines, kernel logistic regression, and *k*-means, can't be easily extended to large collections because their computational complexities are quadratic or even cubic with the size of the training set.

User interface, visualization, and interaction patterns will become more complicated with large quantities of data.

Distributed- and cloud-computing platforms as well as parallel machine-learning and data-mining algorithms will become necessary to make large-scale multimedia analysis run at practical speeds.

The availability of large-scale data might change the way we address the long-standing challenges in multimedia retrieval and mining, which could potentially lead to unexpected breakthroughs.

The articles in this special issue cover a broad variety of research areas that address the challenges introduced by large-scale multimedia collections. We hope that they will serve as a valuable reference for multimedia researchers and developers, but we believe they only reveal the tip of the iceberg in terms of promising directions for research on large-scale multimedia mining. **MM**

Rong Yan is a research scientist at Facebook. His research interests include large-scale machine learning, data mining, social media, multimedia information

retrieval, and computer vision. Yan has a PhD from Carnegie Mellon University's School of Computer Science. Contact him at rongyan@fb.com.

Benoit Huet is an assistant professor in the multimedia information processing group of EURECOM (France). His research interests include computer vision, content-based retrieval, multimedia data mining and indexing (still and/or moving images), and pattern recognition. Huet has a D.Phil in computer science from the University of York, UK, for his research on the topic of object recognition from large databases. Contact him at Benoit Huet Benoit.Huet@eurecom.fr.

Rahul Sukthankar is a senior principal research scientist at Intel Labs Pittsburgh and adjunct research professor in the Robotics Institute at Carnegie Mellon University. His research interests include computer vision and machine learning, particularly in the areas of object recognition, video event detection, and information retrieval. Sukthankar has a PhD in robotics from Carnegie Mellon University. Contact him at rahuls@cs.cmu.edu.