Chapter IV
Cross-Modal Correlation Mining Using Graph Algorithms

Jia-Yu Pan
Carnegie Mellon University, USA

Hyung-Jeong Yang
Chonnam National University, South Korea

Christos Faloutsos
Carnegie Mellon University, USA

Pinar Duygulu
Bilkent University, Turkey

ABSTRACT

Multimedia objects like video clips or captioned images contain data of various modalities such as image, audio, and transcript text. Correlations across different modalities provide information about the multimedia content, and are useful in applications ranging from summarization to semantic captioning. We propose a graph-based method, MAGIC, which represents multimedia data as a graph and can find cross-modal correlations using "random walks with restarts." MAGIC has several desirable properties: (a) it is general and domain-independent; (b) it can detect correlations across any two modalities; (c) it is insensitive to parameter settings; (d) it scales up well for large datasets; (e) it enables novel multimedia applications (e.g., group captioning); and (f) it creates opportunity for applying graph algorithms to multimedia problems. When applied to automatic image captioning, MAGIC finds correlations between text and image and achieves a relative improvement of 58% in captioning accuracy as compared to recent machine learning techniques.

Introduction

Advances in digital technologies make possible the generation and storage of large amount of multimedia objects such as images and video clips. Multimedia content contains rich information in various modalities such as images, audios, video frames, time series, and so forth. However, making

rich multimedia content accessible and useful is not easy. Advanced tools that find characteristic patterns and correlations among multimedia content are required for the effective usage of multimedia databases.

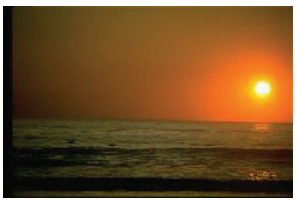
We call a data object whose content is presented in more than one modality a *mixed media* object. For example, a video clip is a mixed media object with image frames, audios, and other information such as transcript text. Another example is a captioned image such as a news picture with an associated description, or a personal photograph annotated with a few keywords (Figure 1). In this chapter, we would use the terms *medium* (plural form *media*) and *modality* interchangeably.

It is common to see correlations among attributes of different modalities on a mixed media object. For instance, a news clip usually contains human speech accompanied with images of static scenes, while a commercial has more dynamic scenes with loud background music (Pan & Faloutsos, 2002). In image archives, caption keywords are chosen such that they describe objects in the images. Similarly, in digital video libraries and entertainment industry, motion picture directors edit sound effects to match the scenes in video frames.

Cross-modal correlations provide helpful hints on exploiting information from different modalities for tasks such as segmentation (Hsu et al., 2004) and indexing (Chang, Manmatha, & Chua, 2005). Also, establishing associations between low-level features and attributes that have semantic meanings may shed light on multimedia understanding. For example, in a collection of captioned images, discovering the correlations between images and caption words could be useful for image annotation, content-based image retrieval, and multimedia understanding.

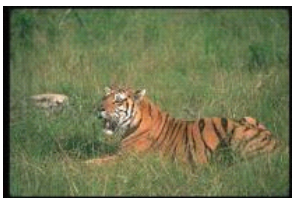
The question that we are interested in is "*Given a collection of mixed media objects, how do we find the correlations across data of various modalities?*" A desirable solution should be able to include all kinds of data modalities, overcome noise in the data, and detect correlations between any subset of modalities available. Moreover, in

Figure 1. Three sample images: (a),(b) are captioned with terms describing the content; (c) is an image to be captioned. (d)(e)(f) show the regions of images (a)(b)(c), respectively.



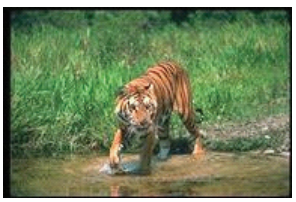
('sea', 'sun', 'sky', 'waves')

(a) Captioned image I_1



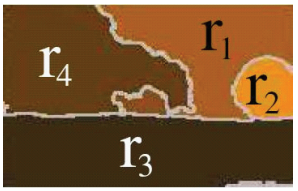
('cat', 'forest', 'grass', 'tiger')

(b) Captioned image I_2



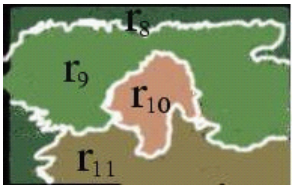
no caption

(c) Image I_3



(d)

(e)



(f)