Chapter VI

Exploiting Captions for Web Data Mining

Neil C. Rowe
U.S. Naval Postgraduate School, USA

ABSTRACT

We survey research on using captions in data mining from the Web. Captions are text that describes some other information (typically, multimedia). Since text is considerably easier to analyze than non-text, a good way to support access to non-text is to index the words of its captions. However, captions vary considerably in form and content on the Web. We discuss the range of syntactic clues (such as HTML tags) and semantic clues (such as particular words). We discuss how to quantify clue strength and combine clues for a consensus. We then discuss the problem of mapping information in captions to information in media objects. While it is hard, classes of mapping schemes are distinguishable, and a segmentation of the media can be matched to a parse of the caption.

INTRODUCTION

Non-text media are an important asset of the World Wide Web. Most of the world prefers to communicate with audio, images, and video rather than written text because these are more natural for the human brain. In addition, much of the world is illiterate and the increasing ubiquity of television and video games is hurting literacy. The Web is the first information technology that permits retrieval of media objects with much the same ease as text. If we could find them, we could get appropriate pictures of orchids or helicopters or quarterbacks in a few seconds without needing to search books or newspapers. Teachers could enliven their lectures with well-chosen images, audio, and video. And a news office could find the perfect picture instead of an adequate one.
Captions are valuable in data mining from the Web. They are text strings that explain or describe other objects, usually non-text or multimedia objects, and they are a form of metadata. Captions help understand and remember media (McAninch, Austin & Derks, 1992, 1993). Captions are especially valuable on the Web because only a small amount of text on Web pages with multimedia (1.2% in a survey of random pages [Rowe, 2002b]) describes the media objects. Thus, standard text browsers, when used to find media matching a particular description often do poorly; if they searched only the captions, they could do much better. Jansen, Goodrum, and Spink (2000) report that 2.65% of all queries in a sample of over 1,000,000 to the Excite search engine were attempting to find images, 0.74% were attempting to find video, and 0.37% were attempting to find audio, so multimedia retrieval was already important in 2000. It will undoubtedly increase in importance as the Internet becomes faster and more people provide Web multimedia resources.

Captions are also valuable because content analysis of media often does not provide the information that users seek. Nontext media do not usually tell when they were created or by whom, what happened before or after, what was happening outside the field of view when they were created, or in what context they were created; and nontext media usually cannot convey linguistic features like quantification, negation, tense, and indirect reference (Cohen, 1992). Furthermore, experiments collecting image-retrieval needs descriptions from users (Armitage & Enser, 1997; Jorgensen, 1998) showed users were rarely concerned with image appearance (e.g., finding a picture with an orange-red circle in the center), but rather, usually with meaning that only captions could provide (e.g., “dawn in the Everglades” or “labor organizing activities”). People seem to have a wide range of tasks for which multimedia retrieval is required, many not needing much understanding of the content of the media (Sutcliffe et al., 1997). This is fortunate because content analysis is often considerably slower and more unreliable than caption analysis because of the often much larger number of bits involved; most useful content analysis requires segmentation, an unreliable and costly process, as well as preprocessing and filtering that is hard to get correct (Flickner et al., 1995; Forsyth, 1999). So finding a caption to a media object simplifies analysis considerably.

But using captions from the Web entails two problems: finding them and understanding them. Finding them is hard because many are not clearly identified: Web page and caption formats and styles vary widely, and the mapping from the world of language to the world of visual or aural concepts is often not straightforward. Nonetheless, tools to address and often solve these problems are available. The somewhat restricted semantics of Web pages and captions can be exploited.
