Abstract

In this age of information, most of the human knowledge resides in some data farms and big servers which are accessed through the internet. Under such a huge information deluge, retrieving the relevant information with minimal effort and time is pivotal. It is not always easy for common users, who lack technical know-how about search systems, to express their information need in a crisp and clear fashion. Consequently, a gap arises between the users' need and the systems interpretation or search results. A large body of research exists in the literature to bridge the gap between what common users need and how systems interpret them. The advent of multimedia data has rendered more complexity to the issue. While plain textual data are relatively easier to store and process, such is not the case with other forms such as image, video, audio, etc. These non-textual data, image, video, audio *etc.* are hard to express in terms of text and not easy to process either. So when it comes to the problem on multimodal information retrieval, one of the primary issues is to reduce the semantic gap between the high-level information need of users and commonly employed low-level features such as image features or audio features and the other one is how to bled or fuse the data of different modalities. We focus on both of these problems regarding combining image as well as text data in our thesis. The reason behind choosing text and image as means of multimodal retrieval is an observation that usually, images are accompanied by text caption along with associated documents describing the image. Existing works on ad-hoc text retrieval and search, show that text query expansion often improves the performance of text retrieval. In contrast, textual query expansion as a form of enhancing image retrieval is a relatively less explored area. Hence, we study the effect of expanding textual query on both image and its associated text retrieval. Our study reveals that judicious expansion of textual query through keyphrase extraction can lead to better results, either in terms of text-retrieval or both image and text-retrieval. To establish this, we use few well-known keyphrase extraction techniques and also propose a novel keyphrase extraction technique to expand the text query. The proposed keyphrase

extraction technique relies on a graph-based model employing mutual information between words occurring within a document as well as across the document collection. We incorporate a topic decomposition based keyphrase extraction technique also in our study to observe the effect of query expansion in multimodal information retrieval. Comparisons are carried out among these keyphrase extraction models using various evaluation measures and the results are reported.

Our next objective is how to blend or fuse these data of different modalities in a prudent way such that the efficiency of multimodal retrieval depends on the way these image and text features are combined. Hence, determining an efficient fusion strategy is crucial. For this purpose, we adopt Fisher-LDA to adjudge the appropriate weights for each modality. This provides us with an intelligent decision-making process favoring the feature set to be infused into the final query.

While dealing with information retrieval, it has a few prerequisites such as the proper processing of data, indexing and organization of the data. Images also need such organization. Images which are not associated with text documents rely on its caption or annotations for text-based or multimodal image retrieval. However, finding a suitable caption for all images on the web or even on storage devices is hard and manual labeling is a cumbersome process. Thus, Automatic Image Annotation (AIA) becomes inevitable in such cases. We propose a novel annotation strategy based on statistical measures and employing a similar graph-based model that we use in our retrieval framework. We argue that instead of annotating every discrete object in an image, we can try and predict what the complete scene in the image depicts, known as "concept". We employ an established community detection algorithm over the concept graph to identify the closest possible annotation for the image. Our proposed strategy is shown to best other state-of-the-art techniques of automatic image annotation.

Other than the image retrieval problem, there is a paradigm shift in object classification where individual feature analysis has given way to multimodal solutions. Thus we find that classifying the images is also a challenge when multiple modalities are associated with it. Like any other multimodal task, in classification task also,

the crucial aspect is how to combine the identified features appropriately so that an optimal result can be obtained with enhanced accuracy. Thus, in our work, we focus on addressing the problem of optimal feature combination for classification task using Hill Climbing. To overcome the shortcomings of Hill Climbing we propose an improved version that increases the classification efficiency and accuracy significantly. Finally, thorough set of experiments on different standard datasets validates our claim that multimodal solutions always perform better than any unimodal one.