Expert Systems With Applications 68 (2017) 81–92

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multimodal Retrieval using Mutual Information based Textual Query Reformulation

Deepanwita Datta*, Shubham Varma[1], Ravindranath Chowdary C., Sanjay K. Singh

Department of Computer Science & Engineering, Indian Institute of Technology (BHU), Varanasi - 221 005, India

ARTICLE INFO

Article history:
Received 18 April 2016
Revised 26 September 2016
Accepted 27 September 2016
Available online 5 October 2016

Keywords:
Multimodal Retrieval
Query Reformulation
Keyphrase Extraction
Mutual Information
Fisher-LDA

ABSTRACT

Multimodal Retrieval is a well-established approach for image retrieval. Usually, images are accompanied by text caption along with associated documents describing the image. Textual query expansion as a form of enhancing image retrieval is a relatively less explored area. In this paper, we first 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 two well-known keyphrase extraction techniques based on *tf-idf* and *KEA*. While query expansion results in increased retrieval efficiency, it is imperative that the expansion be semantically justified. So, we propose a graph-based keyphrase extraction model that captures the *relatedness* between words in terms of both mutual information and relevance feedback. Most of the existing works have stressed on bridging the *semantic gap* by using textual and visual features, either in combination or individually. The way these text and image features are combined determines the efficacy of any retrieval. 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. Our proposed algorithm is shown to supersede the previously mentioned keyphrase extraction algorithms for query expansion significantly. A rigorous set of experiments performed on ImageCLEF-2011 Wikipedia Retrieval task dataset validates our claim that capturing the semantic relation between words through Mutual Information followed by expansion of a textual query using relevance feedback can simultaneously enhance both text and image retrieval.

1. Introduction

Image retrieval is an active research area because of its application in online photo-sharing like Flickr[2], social search (Hoi & Wu (2011)), video-sharing sites like YouTube[3] *etc*. However, due to the ever growing size of the web, simple text-based or visual feature-based retrieval may not be sufficient for optimal search results. Accordingly, there has been a gradual shift towards multimodal retrieval which combines features of both text and image content to achieve better efficiency. This stems from the fact that while textual descriptions are easier to process, they provide very little information about image content. On the other hand, the visual content of images can hardly be described in the text. Thus, an extensive source of information remains underutilized (Depeursinge & Müller, 2010). Also, dealing with visual features is computationally intensive and time-consuming. Hence, conventional approaches (Hoi & Wu, 2011; Zhao & Grosky, 2002) have relied on extracting modalities such as textual and visual features separately and combining them. Visual features are low-level features which can be grouped into (a) Local (interior border, color, texture *etc.*) and (b) Global (shape, contour *etc.*) features (Lisin et al., 2005). Textual features are generally regarded as high-level features. The combination (also called *fusion*) of these features can be done in two ways– *Early Fusion* and *Late Fusion*. In the feature level or early fusion approach, the features extracted from input data are first combined and then sent as input to a single analysis unit. While in the decision level or late fusion approach, the analysis units first provide the local decisions that are obtained based on individual features. The local decisions are then combined using a decision fusion unit to make a fused decision vector that is analyzed further

* Corresponding author.
E-mail addresses: ddatta.rs.cse13@iitbhu.ac.in (D. Datta), shvarma@microsoft.com (S. Varma), rchowdary.cse@iitbhu.ac.in (R. Chowdary C.), sks.cse@iitbhu.ac.in (S.K. Singh).

[1] The author contributed to the work while being an undergraduate student at IIT(BHU), Varanasi and is currently associated with Microsoft India Development Center, Hyderabad.

[2] https://www.flickr.com/.

[3] https://www.youtube.com/?gl=IN.

http://dx.doi.org/10.1016/j.eswa.2016.09.039
0957-4174/

to obtain a final decision (Atrey, Hossain, El Saddik, & Kankanhalli, 2010).

Once the features are extracted and the fusion strategy is determined, the next step usually involves computing proper weights for combining the features (Moulin, Largeron, Ducottet, Géry, & Barat, 2014). Although a significant amount of work has focused on correlating textual and visual information, the *semantic gap* between the high-level information need of users and commonly employed low-level features continues to be a challenge. Semantic representation of documents is not easy to manifest, particularly when non-textual features are involved. Different query formulation techniques such as query-by-example, query-by-sketch, query-by-humming, *etc.* have been suggested in Del Bimbo (1999) to bridge the semantic gap. Whenever queries are involved, *Relevance Feedback*[4] plays a pivotal role (Faro, Giordano, Pino, & Spampinato, 2010; Zellhöfer, 2012). The way a query is formulated often dictates the retrieval results. However, it is not always possible to expect the same level of expertise from an end user as that of a system developer. In such cases, *query reformulation* can prove to be of great use (Hearst, 2009). Among all the query reformulation techniques, query expansion (QE) is highly popular (Mitra, Singhal, & Buckley, 1998). But query expansion itself is not a trivial task and a detailed study is given in Carpineto and Romano (2012). Belkin et al. (2003) observe that query length is positively related to increased search effectiveness. While adding unrelated terms to the query may dilute the focus of the query, appending key concepts (or keyphrases) may enhance the efficiency.

There exists a large number of query expansion techniques in the literature. These methods have been adopted, modified and adapted for various applications ranging from image captioning (Yagcioglu, Erdem, Erdem, & Cakıcı, 2015) to searching geo-data repositories (Hochmair & Fu, 2006). However, their application for image retrieval has been limited (Zhou & Huang, 2002). Even in cases where they have been applied for image retrieval, the semantic relationship between words has not been considered for keyphrase selection. With the help of experiments using two well-established keyword/keyphrase techniques, we show that indeed the performance of an image retrieval can be enhanced. For embedding semantic relationship between terms, we have used Mutual Information and WordNet[5] on a word graph model proposed by us.

In other words, given an initial set of documents (retrieved using a base or user specified query), our proposed keyphrase extraction model generates some succinct set of keyphrases. These keyphrases are then used as a blind feedback to the original query to form an expanded query. To bridge the gap between end-user's actual information need and the retrieved objects, we exploit the relevant part of the narratives (as explained in Section 3.1.1) with the motive of emphasizing what the user really wants. The complete procedure is unsupervised and doesn't require any expertise from end-user, thus, rendering the whole process as an intelligent one.

Our Contributions

Taking a cue from the above concepts, our paper contributes in the following ways.

- Different keyphrase extraction techniques are available in the literature (Hasan & Ng, 2014). Similarly, various studies have been carried out on image retrieval techniques (Lew, Sebe, Djeraba, & Jain, 2006; Li et al., 2016). However, to the best of our knowledge, this is the first attempt which studies the effects of keyphrase extraction based query reformulation on multimodal image retrieval.
- We also hypothesize that inclusion of relevant part of the narratives of a text query may significantly enhance the image retrieval efficiency. This need to stress on relevant part stems from our objective to bring the user closer to her needs in an intelligent fashion.
- We finally propose a new keyphrase extraction approach that tries to capture the semantic relationship between words apart from their co-occurrence. This is actualized with the aid of a word graph formed using mutual information and WordNet. These semantically enriched keyphrases are then used for textual query expansion. In other words, we try to minimize the semantic gap that exists between the image and text by intelligently processing the associated text. The selection of keyphrases from the word graph is made through a *greedy* algorithm.

A comprehensive set of experiments performed on ImageCLEF Wikipedia Retrieval 2011[6] dataset establishes that our proposed textual query expansion method outperforms the other two in both image and accompanying text retrieval.

The remainder of the paper is organized as follows. Section 2 discusses some relevant works in the field of multimodal retrieval. Section 3 delineates the baseline framework. Section 4 presents and compares the various keyphrase extraction approaches in conjunction with query expansion. Our proposed model is also presented in this section. The experimental setup is discussed in Section 6. Results and their associated analysis are presented in Section 7. Section 8 concludes the article while Section 9 lays out some future research directions.

2. Related Work

A vast amount of work can be found in the literature for dealing with various kind of modalities for retrieval. Annotation-based image retrieval (ABIR) simply uses text retrieval techniques on textual annotations of images whereas Content-based image retrieval (CBIR) retrieves images by image contents only. Kilinç and Alpkocak (2011) use image annotations to retrieve images from web pages. In their work, image annotations are modified and enriched by surrounding textual content available. A re-ranking approach is also proposed to improve retrieval performance but image contents are not considered in this work. Such text-based-only or annotation-based image retrieval methods suffer when annotations are missing. In real world, it is hard to expect all images being uniformly annotated. For some, annotations may be missing while others may contain noise. In such cases, content-based image retrieval system comes to the rescue. Yoo, Park, and Jang (2005) propose an content-based expert system using low-level features for image retrieval but they completely disregard any associated text. Yildizer, Balci, Hassan, and Alhajj (2012) propose a fast and efficient CBIR system. Daubechies' wavelets transformation is used to extract feature vectors from the images. Multi-class Support Vector Regression model is applied on those extracted feature vectors for dimension reduction. Finally, the low dimensional feature vectors are classified by a Support Vector Machine (SVM) based classifier which categorizes the entire image databases into different classes. When any query image comes into the system, the categorized image space reduces the searching time and boosts the searching efficiency. The CBIR systems suffers from the drawback that they do not leverage the benefits of associated text. To substantiate the fact that both text and image features facilitates enhanced retrieval performance, current research tend has gradually shifted towards multimodal retrieval.

[4] A comprehensive survey on relevance feedback can be found in (Rocchio, 1971).

[5] https://wordnet.princeton.edu/.

[6] http://medgift.hevs.ch/wikipediaMM/2010-2011/images/.