Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol 2, No.3, 2012

Harvesting Image Databases from The Web

Snehal M. Gaikwad
G.H.Raisoni College of Engg. & Mgmt.,Pune,India
*gaikwad.snehal99@gmail.com
Snehal S. Pathare
G.H.Raisoni College of Engg. & Mgmt.,Pune,India
*snehalpathare4@gmail.com
Trupti A. Jachak
G.H.Raisoni College of Engg. & Mgmt.,Pune,India
*truptijachak311991@gmail.com

Abstract

The research work presented here includes data mining needs and study of their algorithm for various extraction purpose. It also includes work that has been done in the field of harvesting images from web. Here the proposed method is to harvest image databases from web. We can automatically generate a large number of images for a specified object. By applying concept of data mining and the algorithm from data mining which is used for extraction of data or harvesting images. A multimodal approach employing text ,metadata and visual features is used to gather many high-quality images from the web. The modules can be made to find query images by selecting images where nearby text is top ranked by the topic i.e., formation of image clusters then download associate images by using approaches like web search, image search and Google images. Apply re-ranking algorithm and then filtering process to harvest the images.Currently, image search gives a very low precision (only about 4%) and is not used for the harvesting experiments. Since the movements of the technologies are growing rapidly the kinds of work also need to be grown up. This work shows an approach to harvest a large number of images of a particular class automatically and to achieve this with high precision by providing training databases so that a new object model can be learned effortlessly. Many other tools also are available for harvesting images from web .An approach in this paper is original and up to the mark.

Keywords: Legacy code, re-engineering, class diagrams, Aggregation, Association, Attributes

1. Introduction

Now-a-days, mining process is seen as increasingly important tool by modern business into business intelligence giving an information, advantages and also used in marketing, surveillance, fraud detection.

Actual data mining is a process to explore data in search of consistent patterns and /or systematic relationship between variables, and then to validate the findings by applying the detected patterns to new subsets of data.

The focus of this work is intended to harvest a large number of images of particular work automatically and to achieve this with high precision.

2. Objectives

Automatically generating a large number of images for specified object with high precision is arduous manual task. Image search engine apparently provide an effortless route, but currently are limited by poor precision of the returned images and restrictions on the total number of images provided.

The objective of this work is to automatically generate a large number of images for a specified object class. A multimodal approach employing both text, metadata, and visual features is used to gather many high-quality images from the web .Candidate images are obtained by a text-based Web search querying on the object identifier (e.g., the word red rose). The task is then to remove irrelevant images and re-rank the remainder. First, the images are re-ranked based on the text surrounding the image and metadata features. A number of methods are compared for this re-ranking. Second, the top-ranked images are used as (noisy) training data and an SVM visual classifier is learned to improve the ranking further. We investigate the cross-validation procedure to this noisy training data.

This research work is to design and develop a system that generates large number of images for specified object class within high precision by obtaining query images, removing irrelevant images and re-rank the remainder.

3. Proposed Work

This technique prescribed in previous work are used to automatically generate a large number of images for a specified object class. The simple architecture for the proposed system is

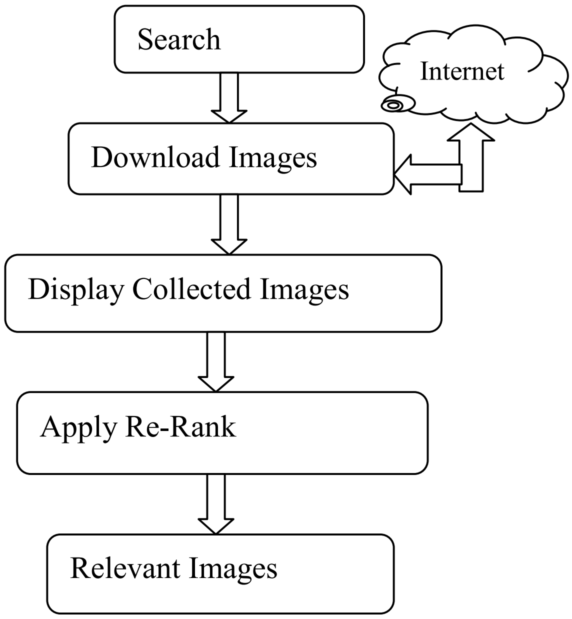

Fig 1: Simple architecture for proposed system

Search Queries: When an image search in search engines, the corresponding images are loaded in that time, meanwhile among them there are also uncategorized images spotted.
Download Images: We compare the three different approaches like Web Search, Image Search, Google Imagesto downloaded images from the web.
Display Collected Images: The results of the applicable images are assembled.
Apply re-rank: The goal is to re-rank the retrieved images. The re-ranking of the assembled images is based on the text and metadata.
Relevant images: To re-rank the filtered images we applied the text+ vision system and hence the relevant images are obtained.

4. System Implementation

The proposed system model illustrates flow of implementation. First, the images are re-ranked based on the text surrounding the image and metadata features. A number of methods are compared for this re-ranking. Second, the top-ranked images are used as(noisy) training data and an SVM visual classifier is learned to improve the ranking further. The principal novelty of the overall method is in combining text/metadata and visual features in order to achieve a completely automatic ranking of the images.

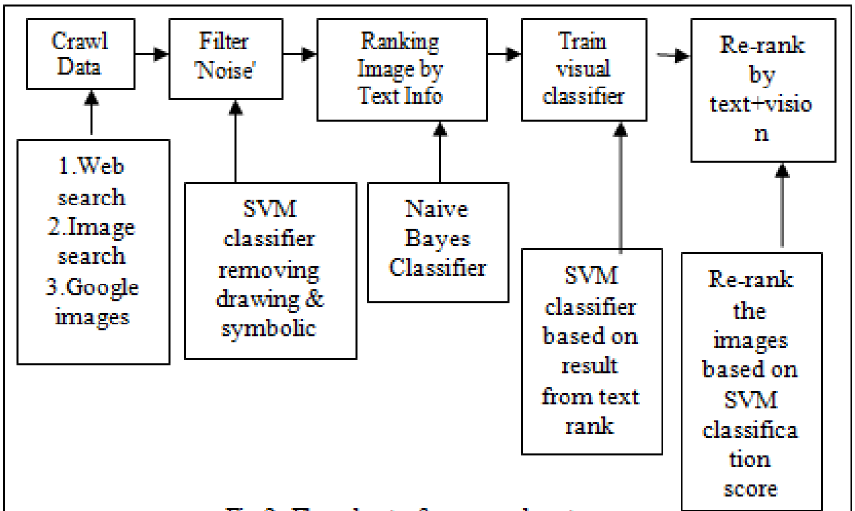
Crawl Data
Filter 'Noise'
Ranking Image by Text Info
Train visual classifier
Re-rank by text+vision
1.Web search 2.Image search 3.Google images
SVM classifier removing drawing & symbolic
Naive Bayes Classifier
SVM classifier based on result from text rank
Re-rank the images based on SVM classification score

Fig 2: Flowchart of proposed system

5. Methodology

5.1 Downloaded candidate images:

This section describes the methods for downloading the initial pool of images from internet.

Crawl Images:

a) ***Web search: S***ubmits the query word to Google web search and all images that are linked within, return web pages which are downloaded(limit 1000pages).

b) ***Google Images:*** Downloaded images are directly returned by Google image search.

c) ***Image Search:*** Each of the returned Google image search is treated as a "seed"-further images are downloaded from the web pages from where the seed image originated.

in-class-good: Images that contain one or many class instances in a clearly visible way.

in-class-ok: Images that show parts of a class instance or obfuscated views of the object due to lighting ,clutter like.

non-class : Images not belonging to in class.

The good and ok sets are further divided into two sub classes,

Abstract: Images that don't look like natural images.

Non-abstract: Images not belonging to the previous class.

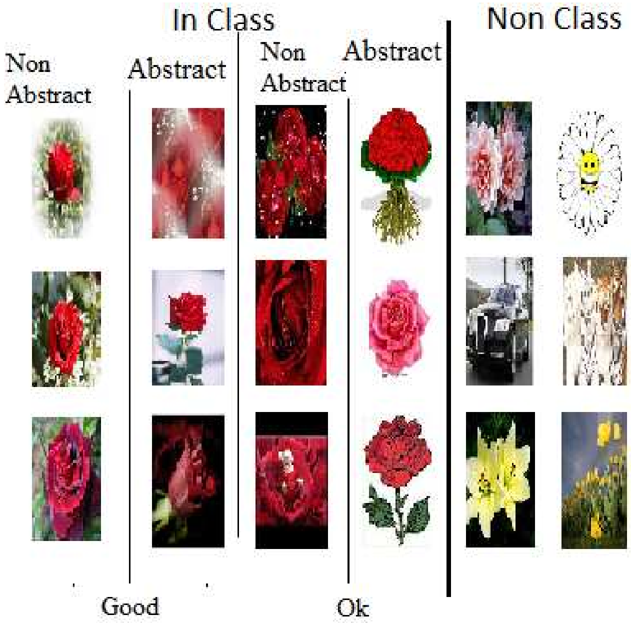

Fig 3: Crawl images

5.2 Removing drawing and symbolic images :

Since we are most likely to have natural image recognition, we would like to remove all abstract images from the downloaded images. Removing abstract images is very challenging for classifiers.

The images contains comics, drawing, sketches etc. We have to remove drawing & abstract images by applying classifiers that separate natural images and drawing images.

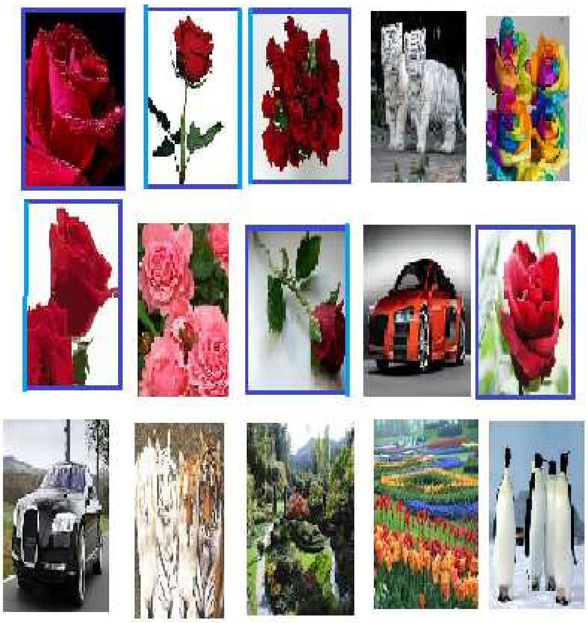

Fig 4: Remove drawing and abstract images

5.3 Apply Re-ranking algorithm

Now we describe the re-ranking of the returned images based on text and metadata alone. The goal is to re-rank the retrieved images. Each feature is treated as binary: "True" if it contains the query words and "false" otherwise. To re-rank images for one particular class (e.g. red rose), we do not employ the whole images for that class. Instead, we train the classifier using all available annotations except the class we want to re-rank. This way, we evaluate performance as a completely automatic class independent image ranker i.e., for any new and unknown class, the images can be ranked without ever using labeled ground-truth knowledge of that class. Rank images using Naive Bayes metadata ranker.

.

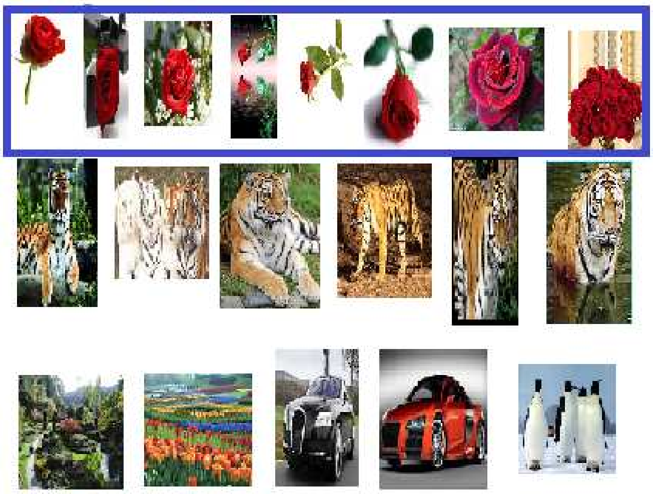

Fig 5: Ranking Images

5.4 *Filtering Process*

The text re-ranker performs well, on average, and significantly improves the precision up to quite a high recall level. To re-ranking the filtered images, we applied the text+ vision system to all images downloaded for one specific class, i.e., the drawings and symbolic images were included.

It is interesting to note that the performance is comparable to the case of filtered images. This means that the learned visual model is strong enough to remove the drawings and symbolic images during the ranking process. Thus, the filtering is only necessary to train the visual classifier and is not required to rank new images,

However, using unfiltered images during training decreases the performance significantly, The main exception here is the airplane class, where training with filtered images is a lot worse than with unfiltered images. In the case of i.e., airplane, the filtering removed 91 good images and the overall precision of the filtered images is quite low, 38.67 percent, which makes the whole process relatively unstable, and therefore can explain the difference.

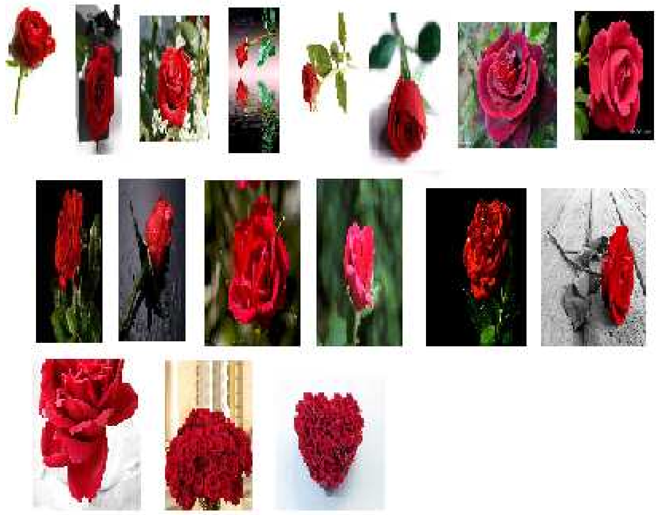

Fig 6: Re-ranking Images

6.Conclusion

Automatically generating a large number of images for a specified object with high precision is not easy task. This paper work provides an approach for automatically generating large number of images of specified object.

7.Acknowledgment

We would like to thank our guide Mrs. Vidya Dhamdhere and head of the department Prof.S. S. Sambare for the encouragement and support that they have extended.r. I would also like to thank the anonymous reviewers who provided helpful feedback on my manuscript.

References

[1] J. Aslam and M. Montague, "Models for Metasearch," Proc. ACMConf. Research and Development in Information Retrieval, pp. 276-284, 2001.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan, "Matching Words and Pictures," J. Machine Learning Research, vol. 3, pp. 1107-1135, Feb. 2003.

[3] T. Berg, "Animals on the Web Data Set," http://www.tamaraberg.com/animalDataset/index.html, 2006.

[4] T. Berg, A. Berg, J. Edwards, M. Mair, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth, "Names and Faces in the News," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

[5] T.L. Berg and D.A. Forsyth, "Animals on the Web," Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

[6] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," J. Machine Learning Research, vol. 3, pp. 993-1022, Jan. 2003.

[7] C.K. Chow and C.N. Liu, "Approximating Discrete Probability Distributions with Dependence Trees," IEEE Information Theory, vol. 14, no. 3, pp. 462-467, May 1968.

[8] B. Collins, J. Deng, K. Li, and L. Fei-Fei, "Towards Scalable Data Set Construction: An Active Learning Approach," Proc. 10th European Conf. Computer Vision, 2008.

[9] N. Dalal and B. Triggs, “Histogram of Oriented Gradients for Human Detection,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 886-893, 2005.

[10] G. Dorko´ and C. Schmid, “Selection of Scale-Invariant Parts for Object Class Recognition,” Proc. Ninth Int’l Conf. Computer Vision, 2003.

[11] R. Fergus, P. Perona, and A. Zisserman, “A Visual Category Filter for Google Images,” Proc. Eighth European Conf. Computer Vision, May 2004.

[1] J. Aslam and M. Montague, “Models for Metasearch,” Proc. ACMConf. Research and Development in Information Retrieval, pp. 276-284, 2001.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan, “Matching Words and Pictures,” J. Machine Learning Research, vol. 3, pp. 1107-1135, Feb. 2003.

[3] T. Berg, “Animals on the Web Data Set,” http://www.tamaraberg.com/animalDataset/index.html, 2006.

[4] T. Berg, A. Berg, J. Edwards, M. Mair, R. White, Y. Teh, E. Learned-Miller, and D. Forsyth, “Names and Faces in the News,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.

[5] T.L. Berg and D.A. Forsyth, “Animals on the Web,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2006.

[6] D. Blei, A. Ng, and M. Jordan, “Latent Dirichlet Allocation,” J. Machine Learning Research, vol. 3, pp. 993-1022, Jan. 2003.

[7] C.K. Chow and C.N. Liu, “Approximating Discrete Probability Distributions with Dependence Trees,” IEEE Information Theory, vol. 14, no. 3, pp. 462-467, May 1968.

[8] B. Collins, J. Deng, K. Li, and L. Fei-Fei, “Towards Scalable Data Set Construction: An Active Learning Approach,” Proc. 10th European Conf. Computer Vision, 2008.

[9] N. Dalal and B. Triggs, “Histogram of Oriented Gradients for Human Detection,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 886-893, 2005.

[10] G. Dorko´ and C. Schmid, “Selection of Scale-Invariant Parts for Object Class Recognition,” Proc. Ninth Int’l Conf. Computer Vision, 2003.

[11] R. Fergus, P. Perona, and A. Zisserman, “A Visual Category Filter for Google Images,” Proc. Eighth European Conf. Computer Vision, May 2004.