EXTRACTING IMAGES FROM THE WEB USING DATA MINING TECHNIQUE

First A. Syed thousif hussain, M.Tech student in Computer Science and Engineering ,. [1]; Second B. N.Kanya, Associate professor of Computer Science and Engineering and Information Technology ,Dr.MGR Educational and Research Institute Chennai, Tamilnadu, INDIA [2];

Abstract

The objective of this work is to generate a large number of images for specified object class. The approach is to employ text, metadata and visual features and to use to gather many high quality images from the web. Candidates images are obtained by text based web search. The web page and the images are downloaded. The task is to remove irrelevant images and to re-rank. First, the images query page is downloaded. Second, it extracts images URL from downloaded page and place it in the database then ranking is done based on text surrounding and metadata features. SVM and Naive bayes classifier algorithm are compared for ranking. The top ranked images are used as training data and an SVM visual classifier is learned to improve re-ranking. The principal idea of the overall method is in combining text or metadata or visual features in order to achieve a completely automatic ranking of images.

Key Terms: Image retrieval, object recognition, computer vision, weakly supervised.

Introduction

Producing a database containing a large number of images and with high precision is still a difficult task. Image search engine apparently provide an effortless route, but currently are limited by poor precision of images and limited on total number of images provided. For example, with Google image search, the precision is low as 32 percent on one of the classes tested here and average 39 percent.

Berg and forsyth overcome the download restriction by starting from web search instead of image search. This search can provide thousands of images. Their method then proceeds in two stages. First, Image clusters for each topic are formed by selecting images where nearby text is top ranked by topic. A user then partitions the clusters into positive and negative for the class. Second, images and the associated text from these clusters are used as exemplars to train a classifier based on voting on visual and text features. The classifier is then used to re-rank the downloaded data set.

Our objective in this work is harvest a large number of images of a particular class and to achieve this with high precision. Our motivation is to provide training databases so that a new object model can be learned easily. The low precision does not allow us to learn a class model from such images. The challenge then is to combine text, Meta data, and visual features in order to achieve the best image re-ranking.

The main contributions are: First, we give a query and that query web page is downloaded. Then we extract the images from that downloaded page and store it in the data bases. Second, the images in the databases can be successfully ranked. The metadata and text attributes on the webpage containing the image provides useful probability and then successfully ranked. The probability is to provide training data for a visual classifier and this classifier delivers a superior reranking to produce by text alone .The class independent text ranker significantly improves this unranked baseline and is itself improved by quite a margin when the vision based ranker is employed. We compared our proposed SVM algorithm to unsupervised methods, concluding that the discriminative approach is better suited for this task, and thus the focus of this work.

The paper is an extended version of "Harvesting Image Databases from the Web". The extension includes a comparison of different text ranking methods, additional features, an investigation of the cross validation to noise in the training data, and a comparison of different topic models.

Related Work

The related work of [1] is to leverage large scale weakly tagged images for computer vision tasks, a novel cross modal tag cleansing and junk image filtering algorithm is developed for cleansing the weakly tagged images and their social tags by integrating both the visual similarity contexts between the images and the semantic similarity contexts their tags. Our algorithm can address the issues of spams, polysemes and synonyms more effectively and determine the relevance between the images and their social tags more precisely, thus it can allow us to create large amounts of training images with more reliable labels by harvesting from large scale weakly tagged images, which can further be used to achieve more effective classifier training.

The related work of [2] present a new approach for modeling multi-modal data sets, focusing on the specific

case of segmented images with associated text. Learning the joint distribution of image regions and words has many applications. It consider in detail predicting words associated with whole images and corresponding to particular image regions. Auto-annotation might help organize and access large collections of images. Region naming is a model of object.

Recognition as a process of translating image regions to words, much as one might translate from one language to another. Learning the relationships between image regions and semantic correlates is an interesting example of multi-modal data mining, particularly because it is typically hard to apply data mining techniques to collections of images. It develop a number of models for the joint distribution of image regions and words, including several which explicitly learn the correspondence between regions and words. We study multi-modal and correspondence extensions to Hofmann's hierarchical clustering/aspect model, a translation model adapted from statistical machine translation , and a multi-modal extension to mixture of latent Dirichlet allocation. All models are assessed using a large collection of annotated images of real.

The related work of [3] present the web holds tremendous potential as a source of training data for visual classification. Web images must be correctly indexed and labeled before this potential can be realized. Accordingly, there has been considerable recent interest in collecting imagery from the web using image search engines to build databases for object and scene recognition research. While search engines can provide rough sets of image data, results are noisy and this leads to problems when training classifiers. They propose a semi-supervised model for automatically collecting clean example imagery from the web. They approach includes both visual and textual web data in a unified framework. Minimal supervision is enabled by the selective use of generative and discriminative elements in a probabilistic model and a novel learning algorithm. It show through experiments that it model discovers good training images from the web with minimal manual work. Classifiers trained using our method significantly outperform analogous baseline approaches on the Caltech-256 dataset.

The related work of [4] proposed a method to improve the results of image search engines on the Internet to satisfy the users who desire to see the relevant images in the first few pages. The results of the text based systems, that use only the accompanied text of the images, are re-ranked by incorporating the visual similarity of the resulting images. It observe that, in general, together with many unrelated ones, the result of text based systems include a subset of correct images, and this set is the largest most similar one compared to other possible subsets. Based on this observation, It present the similarities of all the images in a graph structure, and find the largest densest component of the graph, corresponding to the largest set of most similar subset of images. Then, to re-rank the results, we give higher priority to the images in the densest component, and rank the others based on their similarities to the images in the densest component. The experiments carried out on 10 category of images from promise the success of our method over Google ranking.

Proposed Work

In this proposed paper we are searching a query and then the html page is downloaded to databases. parser is applied to that downloaded page and extracts the image URL form that html page and then the ranking is applied to the images in the databases. Then by using SVM algorithm we are re-ranking the image databases. The architecture diagram is given below:

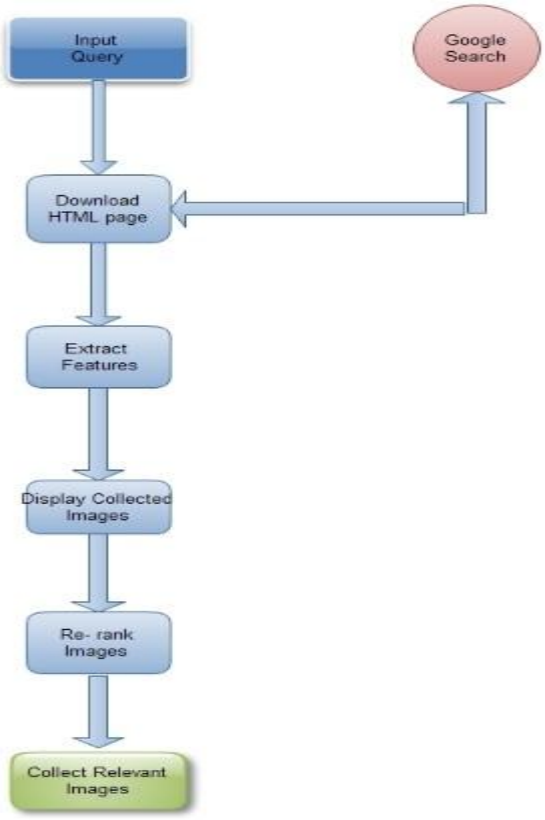

Figure 1: Overall Diagram

The Databases

This section describes the methods for downloading the initial pool of images. We get the data collections by Google image and it search limits the number of returned images to 1,000. Google images, includes only the images directly

returned by Google image search. The query consists of a single word or more specific descriptions. Image smaller than 120 * 120 are removed. The image HTML tag is downloaded with the other metadata such as the image filename.

Table 1 details the statistics for each of the three techniques (web search, Image search and Google Images)

Services	in-class	Non-class	precision
web search	10763	27149	28%
image search	7092	157301	6%
Google Images	5023	75698	42%

TABLE 1: Statistics by source

This low precision is probably due to fact that Google selects many images from web gallery pages which contain images of all sorts. Google is able to select the in-class images from those pages.

Due to great diversity of images available on the internet and because of how we retrieve the images, it is difficult to make general observations on how these databases look. However, it is clear that polysemy affects the returned images. Apart from that, the in-class images occur in almost all variations imaginable. Even though content can clearly be important in re-ranking the images, it will have its limitations due to variety of occurrences of the object.

Removing Drawings and Symbolic Images

We are interested in building databases for natural images recognition, we are would like to remove all abstract images from the downloaded images. However, we have easy way to task for removing drawing and symbolic images.

The removal significantly reduces the number of non-class images, improving the resulting precision of the object class data sets. Filtering out such images can have the aim of removing this type of abstract image from the in-class images.

We train a radial basis function Support Vector Machine on a hand labeled data set. After the initial training, no further user interaction is required. In order to obtain this data set, images were downloaded using Image Search.

The aim was to retrieve many images and then select suitable training images manually. The resulting data set consists of approximately 1,200 drawing and symbolic images and 1,800 non-drawings and symbolic images.

Three visual features are used: a) a color histogram, b) a histogram of L2-norm of the gradient, and c) a histogram of the angles (0…..π) weighted by the L2-norm of the corresponding class.

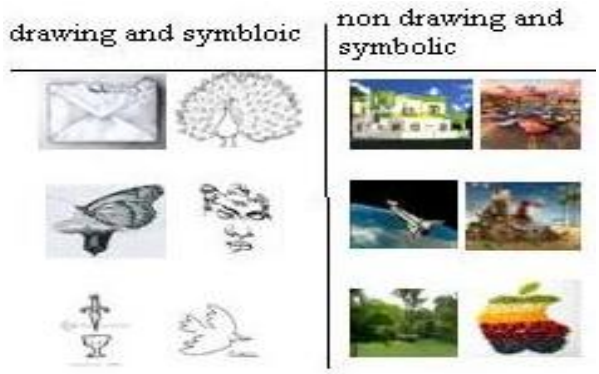

Figure 2: Drawing and Symbolic images

The motivation behind the choice of features is that drawing and symbolic images are characterized by sharp edges in certain orientations. The method achieves around 95 percent classification accuracy on drawing and symbolic images.

The classifier is applied to the entire downloaded image data set to filter drawing and symbolic images before next step. The remaining images are used in over experiment.

Ranking on Textual Features

We now describe the re-ranking of the remaining images based on text and Meta data.

Textual features used by frankel et al. [4] are having seven features from the text and HTML tags on the web page. Textual features context here is defined by the HTML source, not by the rendered page. In the text processing, a standard stop list and porter stemmer are used.

Image Ranking

Using the seven textual features, the goal is to re-rank the retrieve images. Each feature is treated as “true” or “false”. The seven features define a binary feature vector for each image and ranking is based on posterior probability of the image. To re-rank images for one particular class we employ the ground truth data for that class. Using all available annotations except the class we can re-rank the images.

Ranking on Visual Features

The text re-ranking associates a posterior probability with each image as to whether it contained the query class or not. The problem is how to use the information to train a visual classifier that would improve the ranking further. The problem is one of training from noisy data. We can decide which images to use for positive and negative data and to select a validation set in order to optimize the parameters of the classifier.

We use variety of region detectors with common visual features in the data of visual words model framework. All images are first resized to 300 pixels in width. A separate dataset consisting of 100 visual words is learned for each detector using k-means, and these dataset are then combined into a single one of 400 words. Finally, the descriptor of each region is assigned to the dataset. The software for the detectors is obtained.

SVM algorithm

The SVM training algorithm in the ranking of visual features has the following sum.

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C_+ \sum_{i,y-1} \xi_i + C_- \sum_{j,y-1} \xi_i \quad (1)$$

$$\text{Subject to } y_l(w^T \Phi(x_l) + b) \geq 1 - \xi_l \quad (2)$$

$$\xi_l \geq 0 \quad (3)$$

$$l=1,\ldots,(n_+ + n_-) \quad (4)$$

Where x_l are the training vectors and y_l $y_l \in \{1, -1\}$ is the class label. C+ and C- are the classification penalties for the positive and negative images with ξ being the corresponding values.

To implement we use the available SVM light software. The SVM is very sensitive to the parameters, probably due to the huge amount of noise in the data and optimal value does not directly correspond to the ratio of positive to negative images. Finally, the trained SVM is used to re-rank the filtered image set based on the SVM classification. The entire image harvesting algorithm is shown below

1) Giving a query for new class using Google image search.
2) Extract html tags and image url and download the image based on the image url downloaded.
3) Rank images based on text attribute using the SVM.
4) Train visual SVM classifier on text ranked images.
5) Re-rank all images from (4) using the visual classifier

Conclusion

The paper has proposed an automatic algorithm for harvesting the web and gathering hundreds of images of a given query class. Through quantitative evaluation has shown that the proposed algorithm performs similarly to state of art system while performing Google image search and recent techniques that rely on manual intervention.

This paper improves our understating of the further direction could build on top of this understanding as well as ideas and leverage multimodal visual models to extract the different clusters of polysemes meanings Recent work [ ] addresses the only working with few images that are downloaded and shown the result and we are shown in real time environment and the work would be interesting.

Reference

[1] Jianping Fan, Yi Shen, Ning Zhou and Yuli Gao, "Harvesting Large-Scale Weakly-Tagged Image Databases from the Web", Department of Computer Science, UNC-Charlotte, NC28223, USA.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. J. Machine Learning Research, 3:1107–1135, Feb 2003.

[3] Nicholas Morsillo, Christopher Pal and Randal Nelson," Semi-Supervised Learning of Visual Classifiers from Web Images and Text. Department of Computer Science University of Rochester Rochester, NY, D´epartement de g´enie informatique et g´enie logiciel, Ecole Polytechnique De Montreal, Montr´eal, QC, H3C 3A7, Canada.

[4] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image Databases from the Web", Proc. 11th Conf. Computer Vision, 2007.

[5] J. Aslam and M. Montague, "Models for Metasearch". Proc. ACM Conf. Research and Development in Information Retrieval.

[6] K. Barnard, P. Duygulu, N.de Freitas, D. Blei and M. Jordan, "Matching Words and Pictures", J. Machine Learning Research, vol. 3.

[7] R. Fergus, L. Fei-Fei, P. Perona and A. Zisseman, "Learning Object Categories from Google's Image Search", Proc. 10th Int'l Conf. Computer Vision, 2005.

[8] R. Fergus, P. Pernoa and A. Zisserman, "A Visual Category Filter for Google Images", Proc. Eighth European Conf. Computer Vision, May 2004.

[9] C. Frankel, M.J. Swain, and V. Athitsos, "Webseer: An Image Search Engine for the WWW", technical report, Univ. of Chicago, 1997.

[10**]** W.H. Lin, R. Jin and A. Hauptmann, "Web Image Retrieval Re-Ranking with Relevance Model", Proc. IADIS Int'l Conf., 2003.
[11] G. Wang and D. Forsyth, "Object Image Retrieval by Exploiting Online Knowledge Resource", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2008.
[12**]** T.L. Berg and D.A. Forsyth, "Animals on the Web", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2004.
[13] T Joachims, "SVM lisght", http://svmlight.joachims.org/, 2010.
[14] F. Schroff, A, Criminisi and A.Zisserman, "Harvesting Image Databases From The Web", http://www.robots.ox.ac.uk/~vgg/data/mkdb, 2007 .
[15] Carole Bouchard, Jean Frencols Omhover, " A Kansei Based Image Retrieval System Based on The Conjoint Trends Analysis Method.

Biographies

(1)**SYED THOUSIF HUSSAIN** received the B.tech degree in computer science and engineering from The JNTU University, Anantapur, Andhra Pradesh, India in 2010. Currently, He is an M.tech student of Dr M.G.R University, Chennai, Tamil Nadu, India. Syed Thousif Hussain may be reached at thousifsyed@gmail.com, thousif9966@hotmail.com.

(2) **N.KANYA** received the B.S. degree in computer science and engineering from Manonmanian Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India in 1998, and received the M.Sc degree in Information Technology from Alagappa University, Karaikudi, Tamil Nadu, India in 2004, and received the M.tech degree in Computer Science and Engineering from Dr. M.G.R University, Chennai, Tamil Nadu, India in 2007.Currently Pursuing Ph.D in Computer Science and Engineering in the field of Data Mining from Manonmanian Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India. Currently, she is an associate professor of Computer Science And Engineering And Information Technology at Dr. M.G.R University, Chennai, Tamil Nadu, India. Her teaching and research areas include Data Mining, Data Structures, Data Base Technology, Mobile Computing, Bio-infometrices, Software Engineering, Fault Tolerance Systems. She has published 4 Research papers in International Conferences and 10 Research papers in National Conferences. N. Kanya (Associate Professor) may be reached at kanyamtech@yahoo.co.in.