Web Image Annotation Using an Effective Term Weighting

Vundavalli Srinivasarao and Vasudeva Varma

Search and Information Extraction Lab
International Institute of Information Technology
Gachibowli, Hyderabad-32, India
srinivasarao@research.iiit.ac.in, vv@iiit.ac.in

Abstract. The number of images on the World Wide Web has been increasing tremendously. Providing search services for images on the web has been an active research area. Web images are often surrounded by different associated texts like ALT text, surrounding text, image filename, html page title etc. Many popular internet search engines make use of these associated texts while indexing images and give higher importance to the terms present in ALT text. But, a recent study has shown that around half of the images on the web have no ALT text. So, predicting the ALT text of an image in a web page would be of great use in web image retrieval. We propose an approach on top of term co-occurrence approach proposed in the literature to ALT text prediction. Our results show that our approach and the simple term co-occurrence approach produce almost the same results. We analyze both the methods and describe the usage of the methods in different situations. We build an image annotation system on top of our proposed approach and compare the results with the image annotation system built on top of the term co-occurrence approach. Preliminary experiments on a set of 1000 images show that our proposed approach performs well over the simple term co-occurrence approach for web image annotation.

1 Introduction

With the advent of digital devices like digital cameras, camera-enabled mobile phones, the number of images on the World Wide Web is growing rapidly. Providing search services for the web images has been difficult. Traditional image retrieval systems assign annotations to each image manually. Although it is a good methodology to retrieve images through text retrieval technologies, it is gradually becoming impossible to annotate images manually one by one due to the huge and rapid growing number of web images.

Automatic Image Annotation has become more and more important and witnessed rapid development in recent years.

Even the search giant, Google, has attempted to recruit its users to tag random images from its index (see figure 1), by re-framing the process as a collaboration between users with those tags matching between users selected as the labels

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 286–296, 2012.

for the images to improve the quality of Google's image search results[1]. Given that the search giant is using this manual means of image tagging demonstrates the difficulty inherent in the automated image tagging process particularly with regard to scaling those models suggested in the literature to multi-million scale web images and other image libraries.

Fig. 1. Google Image Labeler

A common view is that semantics of web images are well correlated with their associated texts. Because of this, many popular search engines offer web image search based only on the associated texts. ALT text is considered the most important of all associated texts. ALT attribute is used to describe the contents of an image file. It's important for several reasons: ALT attribute is designed to be an alternative text description for images. It represents the semantics of an image as it provides useful information to anyone using the browsers that cannot display images or image display disabled.

Many popular internet search engines like Google Image Search[2] make use of these associated texts while indexing the images and give higher importance to the terms present in ALT text. Google states the importance of ALT text in their official blog[3]: *"As the Googlebot does not see the images directly, we generally concentrate on the information provided in the "ALT" attribute."*

[1] `http://images.google.com/imagelabeler` – Google shut it down in September 2011.

[2] `http://images.google.com/`

[3] `http://googlewebmastercentral.blogspot.com/2007/12/using-alt-attributes-smartly.html`

The ALT attribute has been used in numerous research studies of Web image retrieval. It is given the highest weight in [1]. In [2], the authors consider query terms that occur in ALT text and image names to be *'very relevant'* in ranking images in retrieval. Providing *'a text equivalent for every non-text element'* (for example, by means of the ALT attribute) is a checkpoint in the W3C's Web Content Accessibility Guidelines[4]. The authors of [3] also state the importance of using ALT text. However, a recent study[4] has shown that around half of the images on the web have no ALT text at all. The author collected 1579 images from Yahoo!'s random page service and 3888 images from the Google directory. 47.7% and 49.4% of images respectively had ALT text, of which 26.3% and 27.5% were null. It is clear from this study that most of the web images don't contain ALT text.

[5] proposed a term weighting model based on term co-occurrences to predict the terms in ALT text. One advantage of the approach is that it can be extended to any image dataset with associated texts. In this paper, we combine the above term weighting model with the natural language processing applications. We build an image annotation system on top of the above term weighting model and our proposed model.

The reminder of the paper is organized as follows. Section 2 gives an overview of related work. In Section 3, we describe our proposed approach to term weighting using term co-occurrences and natural language processing applications. We describe the dataset, evaluate our system and prove the usefulness of it in Section 4. We summarize the paper and give an account of our future directions in Section 5.

2 Related Work

There has been plenty of work done in Automatic Image Annotation. Some of the early approaches[6,7] to image annotation are closely related to image classification. Images are assigned a set of sample descriptions(predefined categories) such as people, landscape, indoor, outdoor, animal. Instead of focusing on the annotation task, they focus more on image processing and feature selection.

Co-occurrence Model[8], Translation Model[9], Latent Dirichlet Allocation Model[10], Cross-Media Relevance Model[11], etc infer the correlations or joint probabilities between images and annotation keywords. Other works like linguistic indexing[12], and Multi-instanced learning[13] try to associate keywords (concepts) with images by learning classifiers. To develop accurate image annotation models, some manualy labeled data is required. Most of the approaches mentioned above have been developed and tested almost exclusively on the Corel[5] database. The latter contains 600 CDROMs, each containing about 100 images representing the same topic or concept, e.g., people, landscape, male. Each topic is associated with keywords and these are assumed to also describe the images under this topic.

[4] Web content accessibility guidelines 1.0. Retrieved 26 August, 2005 from `http://www.w3.org/TR/WAI-WEBCONTENT/`

[5] `http://www.corel.com/`

[14] demonstrates some of the disadvantages of data-sets like Corel set for effective automatic image annotation. It is unlikely that models trained on Corel database will perform well on other image collections. Web images differ from general images in that they are associated by plentiful of texts. Various approaches[15,16] have been employed to improve the performance of web image annotation based on associated texts. Our work differs from previous work in that our approach is evaluated on a large number of images and works well for any image dataset with associated texts.

3 Term Weighting Model

In this section, we propose a term[6] weighting model which is a combination of the model proposed in [5] and natural language processing applications. We will give a brief overview of the model proposed in [5]. The model computes the term weights based on the term co-occurrences in image associated texts to predict the terms in ALT text. A term is said to be important if it occurs in many associated texts and co-occurs with many other terms present in different associated texts.

For each image, we calculate term weights using the following equation.

$$w(t) = (\frac{\sum_i s(t,t_i)}{N})(Imp(t)) \tag{1}$$

$s(t,t_i)$, the similarity between two terms t and t_i, is calculated using Jaccard Similarity as follows:

$$s(t,t_i) = \frac{|S_t \cap S_{t_i}|}{|S_t \cup S_{t_i}|} \tag{2}$$

$S_t \cap S_{t_i}$ is the set of associated texts in which both t and t_i occur, $|S_t \cup S_{t_i}|$ can be calculated as $|S_t| + |S_{t_i}| - |S_t \cap S_{t_i}|$, and S_t is the set of associated texts which contain the term t. N is the total number of unique terms in all associated texts. $Imp(t)$ is the importance of a term which is calculated as follows:

$$Imp(t) = \frac{\sum_i boost(a_i)}{|A|} \tag{3}$$

$boost(a_i)$ is the boost of the associated text a_i which contains the term t and A is the set of associated texts which contain the term t. The extracted associated texts are assigned a boost based on the heuristic of importance(image caption, HTML title, image filename, anchor text, source page url, surrounding text in that order). Value of boost for each associated text is given based on the importance of the associated text as stated above. Once the weight for each term has been computed, the terms are ranked in descending order based on term weights and top k terms are selected as terms in ALT text.

[6] We use *term and word interchangeably in this paper.*

Based on the assumption that noun phrases(NP) are the best lexical category to describe the images[17], we extract noun phrases from the associated texts and consider the terms present in noun phrases as candidate terms for the ALT text. We then considered the terms present in both noun phrases and verb phrases as candidate terms for ALT text. Next, we combined the above approaches with the term co-occurrence approach where the terms in the extracted noun phrases and verb phrases are given more importance. We use OpenNLP tool[7] to extract noun phrases and verb phrases.

We describe the evaluation procedure and present the results of the approaches in the following section.

4 Evaluation

In this section, we present the evaluation procedure of our approach. We briefly describe the data collection and preprocessing steps, present the evaluation procedure and finally results are discussed.

4.1 Data Collection and Preprocessing

A crawler is used to collect images from many websites. Images like banners, icons, navigation buttons etc, which are not so useful are not considered. The web documents are preprocessed to extract associated texts so that the images can be indexed and retrieved with these text features. The associated texts we considered are extracted from ALT attribute, HTML page title, image filename, source page url, anchor text, image caption and surrounding text.

We used Guardian[8], Telegraph[9] and Reuters[10] as the source urls and collected a total of 200000 images which have ALT text. We selected these news websites because the ALT text provided in them is accurate and is very useful for evaluation. The pages in which the images are present, cover a wide range of topics including technology, sports, national and international news, etc. Stopwords are removed and stemming is used to filter the duplicate words from the extracted textual information.

We compare our results with the term co-occurrence approach proposed in [5].

4.2 Evaluation Procedure

In order to evaluate the effectiveness of our method, we compare the predicted terms produced by our approach against the terms extracted from ALT attribute of an image in the corresponding web page.

[7] http://incubator.apache.org/opennlp/
[8] http://www.guardian.co.uk/
[9] http://www.telegraph.co.uk/
[10] http://www.reuters.com/

We present results using the top 5, 10, and 15 words. We adopt the recall, precision and F-measures to evaluate the performance in our experiments. If P_t is the set of terms predicted by our approach and A_t is the set of terms in ALT text, then in our task, we calculate precision, recall and F-measure as follows:

$$precision = \frac{Number\ of\ common\ terms\ between\ P_t\ and\ A_t}{Total\ number\ of\ terms\ in\ P_t} \tag{4}$$

$$recall = \frac{Number\ of\ common\ terms\ between\ P_t\ and\ A_t}{Total\ number\ of\ terms\ in\ A_t} \tag{5}$$

$$F - Measure = \frac{2 * precision * recall}{precision + recall} \tag{6}$$

4.3 Analysis

The approach, NP, considers the terms in noun phrases as candidate terms for ALT text. Similarly, NP + VP, considers the terms in both noun phrases and verb phrases as candidate terms for ALT text.

NP + Term co-occurrence is our proposed term co-occurrence approach in which we give more boost to the terms in noun phrases. Similarly, NP+VP+Term co-occurrence is our proposed term co-occurrence approach in which we give more weights to the terms present in noun phrases and verb phrases.

As we can observe from the results in tables 1 to 3, both term co-occurrence approach and term co-occurrence approach combined with NP+VP give better results compared to other approaches and they give almost the same results.

Table 1. Comparison of approaches for top 5 predicted terms for 200000 images

Approach	**Precision@5**	**Recall@5**	**F-Measure@5**
Term co-occurrence	53.19	41.95	**46.91**
NP	40.78	33.59	36.83
NP + VP	41.67	34.78	37.91
NP + Term co-occurrence	48.49	39.24	43.38
NP + VP + Term co-occurrence	49.79	41.46	45.24

For ex: Consider the figures 2 to 5.

Figure 2 is the image of a doll bearing the faces of Russian leader Vladimir Putin and Dmitry Medvedev. The ALT text of the image is *"Dmitry Medvedev: Vladimir Putin is more popular than me"*. The term 'popular' is predicted by the term co-occurrence approach where as it is not predicted by any of other approaches.

Figure 3 is the image of Jose Mourinho. The ALT text of the image is *"Jose Mourinho handed two-match ban for Super Cup eye poke"*. The term 'poke' is predicted by the term co-occurrence approach where as it is not predicted by any of other approaches.

Table 2. Comparison of approaches for top 10 predicted terms for 200000 images

Approach	Precision@10	Recall@10	F-Measure@10
Term co-occurrence	44.11	60.87	51.15
NP	36.87	53.84	43.76
NP + VP	38.13	55.81	45.30
NP + Term co-occurrence	39.07	56.14	46.07
NP + VP + Term co-occurrence	43.60	61.86	51.15

Table 3. Comparison of approaches for top 15 predicted terms for 200000 images

Approach	Precision@15	Recall@15	F-Measure@15
Term co-occurrence	37.96	72.98	**49.94**
NP	32.14	61.42	42.19
NP + VP	32.73	65.87	43.73
NP + Term co-occurrence	31.79	62.41	42.12
NP + VP + Term co-occurrence	37.35	69.25	48.52

Figure 4 is the image of Nicolas Sarkozy and Angela Merkel in a meeting. The terms Angela and Merkel are very relevant to the image. But they are not predicted by term co-occurrence approach where as they are predicted by term co-occurrence approach combined with NP+VP.

Figure 5 is the image of Casey Stoner at Italian MotoGP. The term MotoGP is very relevant to the image. But it is not predicted by the term co-occurrence approach, where as it is predicted by term co-occurrence approach combined with NP+VP.

There are a few such cases where we missed such important entities from the predicted terms with the term co-occurrence approach. And also some of the terms in the alternate text are not predicted by approaches other than the term co-occurrence approach.

We build an image annotation system on top of the term co-occurrence approach and compared it with the image annotation system built on top of the term co-occurrence approach combined with NP+VP. For the evaluation of image annotation system, we select a subset of 1000 images from the original dataset. Human annotators are chosen to manually assign tags to the images.

Given an image, the following points are taken into account while annotating the image.

- The subject who annotates the image, assigns the keywords which he/she thinks are relevant to the image, without taking a look at the web page which contains the image.
- Once he/she assigns the keywords to the image, he/she visits the web page which contains the image and refines the annotations using the terms extracted from the associated text.

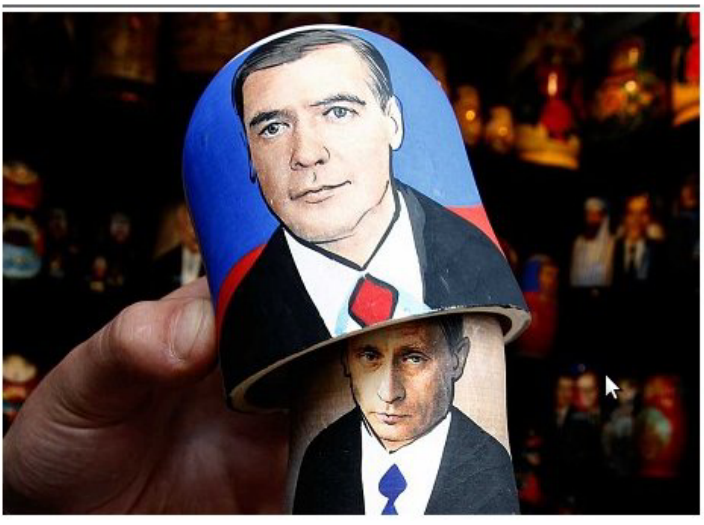

Fig. 2. A doll bearing the faces of Russian leader Vladimir Putin and Dmitry Medvedev

Fig. 3. Image of Jose Mourinho

As we can see from the results, tables 4 to 6, of the image annotation systems, term co-occurrence+NP+VP gives better results over simple term co-occurrence. For the task of predicting terms in ALT text, both simple term co-occurrence and term co-occurrence + NP + VP work well. However, we prefer simple term co-

Fig. 4. Image of Nicolas Sarkozy and Angela Merkel

Fig. 5. Image of Casey Stoner at Italian MotoGP

occurrence approach over term co-occurrence+NP+VP as the former is language independent.

Table 4. Comparison of annotation systems for top 5 annotations

Approach	Precision@5	Recall@5	F-Measure@5
Term co-occurrence	53.97	50.26	52.04
NP + VP + Term co-occurrence	55.83	52.92	54.34

Table 5. Comparison of annotation systems for top 10 annotations

Approach	Precision@10	Recall@10	F-Measure@10
Term co-occurrence	33.95	64.26	44.43
NP + VP + Term co-occurrence	36.04	67.76	47.05

Table 6. Comparison of annotation systems for top 15 annotations

Approach	Precision@15	Recall@15	F-Measure@15
Term co-occurrence	25.27	70.44	37.20
NP + VP + Term co-occurrence	26.80	75.14	39.51

5 Conclusions and Future Work

In this chapter, we presented a term weighting approach that makes use of term co-occurrences in associated texts and predicts terms occurring in ALT text of an image. We compared the performance of our approach against a few baseline approaches which use term frequency, document frequency, terms in noun phrases and terms in verb phrases respectively. Experiments on a large number of images showed that our model is able to achieve a good performance for the prediction task. We built image annotation systems on top of the above approaches and found out that the term co-occurrence approach in combination with noun phrases and verb phrases performs better than the term co-occurrence approach. The simple term co-occurrence approach for the prediction of terms in alt text is language independent and is preferable over term co-occurrence combined with noun phrases and verb phrases even though both of them produce almost same results. However, the later performs well for the task of web image annotation.

For web image annotation task, we would like to experiment on more number of images and come to a conclusion that the term co-occurrence approach combined with NP+VP works better than the simple term co-occurrence approach for the annotation task.

References

1. Cascia, M.L., Sethi, S., Sclaroff, S.: Combining textual and visual cues for content-based image retrieval on the world wide web. In: IEEE Workshop on Content-based Access of Image and Video Libraries, pp. 24–28 (1998)
2. Mukherjea, S., Hirata, K., Hara, Y.: Amore: A world wide web image retrieval engine. World Wide Web 2, 115–132 (1999)
3. Petrie, H., Harrison, C., Dev, S.: Describing images on the web: a survey of current practice and prospects for the future. In: Proceedings of Human Computer Interaction International, HCII 2005 (2005)
4. Craven, T.C.: Some features of alt texts associated with images in web pages. Information Research 11 (2006)
5. Srinivasarao, V., Pingali, P., Varma, V.: Effective Term Weighting in Alt Text Prediction for Web Image Retrieval. In: Du, X., Fan, W., Wang, J., Peng, Z., Sharaf, M.A. (eds.) APWeb 2011. LNCS, vol. 6612, pp. 237–244. Springer, Heidelberg (2011)
6. Vailaya, A., Figueiredo, M.A.T., Jain, A.K., Zhang, H.J.: Image classification for content-based indexing. IEEE Transactions on Image Processing 10, 117–130 (2001)
7. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1349–1380 (2000)
8. Hironobu, Y.M., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. Boltzmann machines, Neural Networks 4 (1999)
9. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
10. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: SIGIR 2003: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 127–134 (2003)
11. Jeon, J., Lavrenko, V., Manmatha, R., Callan, J., Cormack, G., Clarke, C., Hawking, D., Smeaton, A.: Automatic image annotation and retrieval using cross-media relevance models. SIGIR Forum, 119–126 (2003)
12. Jia, L., Wang, Z.J.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1075–1088 (2003)
13. Yang, C., Dong, M.: Region-based image annotation using asymmetrical support vector machine-based multi-instance learning. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2 (2006)
14. Tang, J., Lewis, P.: A study of quality issues for image auto-annotation with the corel data-set. IEEE Transactions on Circuits and Systems for Video Technology 1, 384–389 (2007)
15. Rui, X., Li, M., Li, Z., Ma, W.Y., Yu, N.: Bipartite graph reinforcement model for web image annotation. ACM Multimedia, 585–594 (2007)
16. Shen, H.T., Ooi, B.C., Tan, K.L.: Giving meaning to www images. ACM Multimedia, 39–47 (2000)
17. Kuo, C.H., Chou, T.C., Tsao, N.L., Lan, Y.H.: Canfind: A semantic image indexing and retrieval system. In: ISCAS (3), pp. 644–647 (2003)