IMU @ ImageCLEF 2012

Xueliang Yan[⋆], Wei Wu[⋆], Guanglai Gao[⋆⋆], and Qianqian Lu

College of Computer Science
Inner Mongolia University
010021 Hohhot, China
{csyxl,cswuwei,csggl,cslqq}@imu.edu.cn

Abstract. Inner Mongolia University have participated the Visual concept detection, annotation, and retrieval using Flickr photos task of ImageCLEF for the first time in 2012. We have conducted experiments and submitted results for both the Concept Annotation and the Concept-based Retrieval subtasks. This paper describes the methods we have adopted and the analysis of the results for the two subtasks. We focus our attention mainly on the user's tag since we believe that user annotation provides strong semantic information which can be used to accurately determine the presence or absence of each concept and the relevance level between the images and queries. For the Concept Annotation subtask, we use only a simple statistical method that scores the confidence of the presence of each concept by the maximum conditional probability of the concept between the different given tags. For the Concept-based Retrieval task, we adopted the language modeling approach which has been widely used in text information retrieval field. Official evaluations show that the performance of our method is competitive. We rank in the middle of the pack for the Concept Annotation subtask with the best run's MiAP equal 0.2441. For the Concept-based Retrieval subtask, we rank at the top with the best run's MnAP equal 0.0933. Beside the main submissions, we also submit two visual runs, although no very good, with the MiAP for Concept Annotation is 0.0819 and the MnAP for Concept Retrieval is 0.0045. As a whole, the results confirm that although the methods we have adopted are simple, the performances we have achieved are satisfied.

Keywords: Concept Annotation, Concept Retrieval, User Tags, Wiki Expansion, Maximum Conditional Probability, Language Modeling, Bag-of-Visual Words

1 Introduction

In this paper we will describe the approaches we have adopted to accomplish the "Visual concept detection, annotation, and retrieval using Flickr photos"

[⋆] The first two authors have equal contribution in this work and both are equally considered as 'first author'.

[⋆⋆] Corresponding Author

task for ImageCLEF 2012. We have participated both the Concept Annotation subtask and the Concept-based Retrieval subtask and submitted 15 runs in total. The official evaluation shows that we rank in the middle pack in the Concept Annotation subtask and get the highest rank in the Concept-based Retrieval subtask[1]. We base our methods mainly on User Tags, both for the Concept Annotation and for the Concept-based Retrieval subtask. The main reason for choosing User Tag as feature is that we believe that users generally annotate an image with the words that have strong relationships to its meaning. This semantic information can then be used to determine the contents of the image accurately. In all, we used statistical methods to address both the two subtasks. For the annotation subtask, we use the training set to estimate a conditional probability distribution and use the probability of the most supportive tag as the confidence of the presence of a concept in an image. For the retrieval subtask, we construct language models for the tags of each image and take the probability of the query being generated by the tags model as ranking score. Beside the main methods, we also test some visual feature based method.

The rest of the paper is organized as follows: in section 2 we firstly discuss the method we have used to accomplish the Concept Annotation subtask, including the experiments and the results achieved. Then in section 3, we similarly discuss the method for Concept Retrieval, again, experiments and result analysis are included. Finally, we conclude our work and shed lights on the future work in section 4 and 5 respectively.

2 Concept-based Annotation

2.1 User Tag based Method for Annotation

We have adopted a very simple statistical method for the Concept-based Annotation subtask here. That is: we just calculate the conditional probability of the presence of the concept given the tags of the image and then taking the maximum probability among the different tags as the confidence. This can be shown by (1):

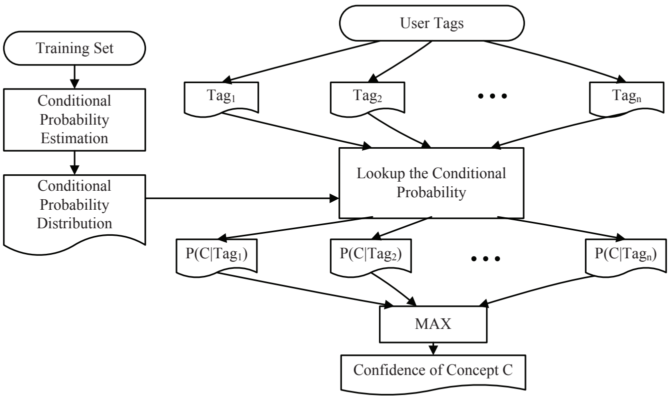
$$Confidence(C) = \max_{Tag} P(C|Tag) \ , \tag{1}$$

where C denotes the concept being considered; $Confidence(C)$ is the confidence of the presence of concept C.

The flowchart of our Tag-based Concept Annotation system is illustrated in Fig. 1.

We use the training set which has been released by CLEF organization to estimate the conditional probability distribution. Since the number of tags for each image is relatively small (6.34 on average according to our statistics), we think that it's reasonable to expand them first[2]. We use the Official INEX 09 collection[3] which contains about 2640000 articles to perform User Tag expansion. [1] More specifically, we first use the tags for each image as query and the INEX

[1] http://www.mpi-inf.mpg.de/departments/d5/software/inex/



User Tags
Training Set
Conditional Probability Estimation
Conditional Probability Distribution
Tag1
Tag2
• • •
Tagn
Lookup the Conditional Probability
P(C|Tag1)
P(C|Tag2)
• • •
P(C|Tagn)
MAX
Confidence of Concept C

Fig. 1. Flowchart of the Tag-based Concept Annotation System

Table 1. Collections for Tag-based Concept Annotation

Collection Name	**Collection Scale**
Training set for Concept Annotation of ImageCLEF2012	15000 images
Testing set for Concept Annotation of ImageCLEF2012	10000 images
Wiki INEX09	2640000 articles

corpora as document collection to perform retrieval and then extend the tags by using the top 15 words among the returned top 200 documents. We test our annotation approach on the officially released testing set, which contains 10000 images in total. The collections we used for our Tag-basd annotation approach are listed in Tab. 1.

We use Indri[4] in the Lemur Toolkit[2] to perform retrieval. Indri is a search engine which has been widely used in the Information Retrieval field. We will mention it again in the Concept-based Retrieval section (section 3). The parameters set in Indri for the Tag-based Annotation are listed in Tab. 2.

The experimental results are summarized in Tab. 3. In Tab. 3,Max_CondProb denotes Maximum Conditional Probability; WE_Train_Max_CondtProb means Maximum Conditional Probability with tag expansion for Training Set by Wiki INEX09; WE_Train_Test_Max_CondProb is the Maximum Conditional Probability method with tag expansion for both training and testing set by Wiki INEX 09; WE_Train_Test_Norm means Maximum Conditional Probability with tag expansion for both training and testing set by Wiki INEX 09 and scores are normalized to the range of 0 to 1.

Comparing the first two results Max_CondProb and WE_Train_Max_CondProb in Tab. 3, we can see that it really reaps the benefit of the Wiki Expansion greatly (more than 1.0 percent increase for the MiAP matric).

[2] `http://www.lemurproject.org`

Table 2. Parameters set in Indri for Tag-based Annotation

Parameter_Name	**Parameter_Value**
Score Function	KL-divergence
Smoothing Method	Dirichlet with =2500
#doc Returned	1000 per query
Stop Words	418 for query
Feedback_for_Tag_Expansion	fbTerm=15, fbDocs=200

Table 3. Evaluation Results for Concept Annotation with User Tags as Features

Result_Name	**MiAP**	**GMiAP**	**F-ex**
Max_CondProb	0.2241	0.1698	0.4128
WE_Train_Max_CondProb	0.2368	0.1825	0.4685
WE_Train_Test_Max_CondProb	0.2174	0.1665	0.4535
WE_Train_Test_Norm_Max_CondProb	0.2441	0.1917	0.4535

The huge drop from the performance of WE_Train_Max_CondProb to that of WE_Train_Test_Max_CondProb indicates that it's not reasonable to expand the tags for the images to be annotated. We can explain this as that the expanded tags are more likely to drift away to the concepts that are not in fact exist in the given image.

Intuition tells us that every image should contain some concepts in general, otherwise it will not be chosen for sharing. This means that even though the absolute probability of a concept may be low, we should still have great confidence that the concept is present if it has a relatively higher probability than the other concepts. This assumption can be confirmed by comparing WE_Train_Test_Max_CondProb to the run with the probability value normalized across the concepts (WE_Train_Test_Norm_Max_CondProb). The great improvement in performance shows that normalization plays an important role in transforming the conditional probability to the confidence of the concept.

2.2 Visual Feature based Method for Annotation

Extraction of Visual Features We extracted three features that are mostly considered in the literatures we found (i.e., Color Histograms, Fuzzy-Color-and-Texture-Histogram (FCTH) and Bag-of-Visual Words).

Color Histograms are among the most basic approaches and are widely used in image retrieval. The color space is partitioned and for each partition the pixels with their color within this range are counted, resulting in a representation of the relative frequencies of the colors. We use the RGB color space for the histograms[5]. And we use the Jensen-Shannon divergence (JSD) as shown in

(2) to compute the distance:

$$d_{JSD}(H,H^{'}) = \sum_{m=1}^{M} \left[H_m \log \frac{2H_m}{H_m + H^{'}_m} + H^{'}_m \log \frac{2H^{'}_m}{H_m + H^{'}_m} \right] \quad , \tag{2}$$

where H and H' are the histograms to be compared.

Fuzzy Color and Texture Histogram (FCTH) is appropriate for accurately retrieving images even in distortion cases such as deformations, noise and smoothing. FCTH is a low level descriptor that contains both quantized histogram color and texture information[6]. For the measurement of the distance of this feature between the images, we use Tanimoto coefficient as shown by (3):

$$T_{ij} = t(x_i, x_j) = \frac{x_i^T x_j}{x_i^T x_i + x_j^T x_j - x_i^T x_j} \quad . \tag{3}$$

We extracted the SIFT local features from harris-laplace region of interest detection. Each of these features is represented as a Bag-of-Visual Word. The visual words vocabulary is generated by adopting the K-means clustering algorithm on the features of the training set, which is implemented in the LIRE Toolkit[3][7]. In our experiment, we take 10000 as the size of the Visual Word Vocabulary and adopt the same d_{JSD} in (2) as the distance masure for clusting.

The three features described above are then used independently for k-NN classifier.

Classification Firstly, we use visual features mentioned above to build classifier. We use distance-weighted k-nearest neighbour (k-NN) approach to build our classifiers[8]. For each concept, we selected some positive images and a number of negative images. The distances, for feature f, from the test image T_i to each of the k nearest positive or negative images are determined. Then we computed the similarity between the test image T_i and Concept C as (4):

$$Sim_f(C, T_i) = \frac{\sum_{p \in P} (dist_f(T_i, p) + \varepsilon)^{-1}}{\sum_{n \in N} (dist_f(T_i, n) + \varepsilon)^{-1} + \varepsilon} \quad , \tag{4}$$

where P and N are the k-nearest positive and negative images for each concept and satisfy $|Q| + |N| = k$ and ε is a small positive number to avoid division by zero.

However, the experiment results on the training set are not very good, so we submitted only one visual feature based run (Bag-of-Visual Words).

Tab. 4 lists the official evaluation result of our visual submission, which again confirms that by now our visual based method is not good.

[3] http://www.semanticmetadata.net/lire/

Table 4. Evaluation Results for Concept Annotation with Visual Words as Features

Result_name	MiAP	GMiAP	F-ex
BoV_Annotation	0.0819	0.0387	0.0429

3 Concept Retrieval

3.1 Language Models for User Tag Retrieval

Language modeling is a formal probabilistic framework that has been widely used in the text retrieval field. The language modeling approach to text retrieval is to model the idea that a document is a good match to a query if the document model is likely to generate the query[9]. Formally, we want to estimate a model M_d for each document d and rank the documents for a query Q according to the probability of Q being generated by M_d. When further assuming beg-of-words modeling, we get (5):

$$\log p\left(Q|M_d\right) = \sum_{i=1}^{m} \log p\left(Q_i|M_d\right) \quad . \tag{5}$$

Observed that, in the official image collection, if an image is relevant to a query, its tags annotated by the user are more likely to occur in the query, we guess that the Concept Retrieval task can be addressed by building language models for the image tags and retrieving the images by ranking them according to the likelihood of generating the query (6):

$$\log p\left(Q|M_{ImageTags}\right) = \sum_{i=1}^{m} \log p\left(Q_i|M_{ImageTags}\right) \quad . \tag{6}$$

To test our assumption, we use the Indri Search Engine in the Lemur Toolkit which we have mentioned in section 2 to perform retrieval. Indri is a search engine that implements the language modeling approach under the Bayes Inference Network Framework [10]. The Tag models are smoothed by the Dirichlet Smoothing[11] method. For each topic, we retrieve 1000 images. A total of 418 stop words from the standard InQuery[12] stoplist are removed from the queries. Pseudo Relevance Feedback (PRF)[9] is adopted for both Tag Expansion and Concept Retrieval with fbTerms=15 and fbDocs=200. The detailed parameters set in Indri for Tag-based Retrieval are listed in Tab. 5.

In Tab. 6, we list the collections used to perform our Tag-based retrieval experiments and illustrate the flowchart of our system in Fig. 2.

In our experiments, we have tested different methods to form the queries for language modeling based retrieval and the results we have achieved are listed in Tab. 7 and Tab. 8.

In Tab. 7, Title means that the queries are formulated by just using the terms in the <title> field of the topic file queries.xml which has been officially released in

Table 5. Parameters set in Indri for Tag-based Retrieval

Parameter_Name	**Parameter_Value**
Score Function	KL-divergence
Smoothing Method	Dirichlet with =2500
#Doc Returned	1000 per query
Stop Words	418 for query
Feedback_for_Tag_Expansion	fbTerm=15, fbDocs=200
Feedback_for_Retrieval	fbTerm=15, fbDocs=200

Table 6. Collections for Tag-based Concept Retrieval

Collection Name	**Collection Scale**
Testing set for Concept Retrieval of ImageCLEF2012	200000 images
Wiki INEX09	2640000 articles

the test set. Image(all_3) denotes that the queries are constructed by using all the three images' tags, i.e., the <image> fields in the topic file; Image(first_not_null) refers to the query construction method that only the first image's tags in the topic file are used. If the first image's tag set is empty or the tags cannot return any result, use the next one's tags, and so forth. Title&Image(all_3) means that the queries are constructed by combining the title portion and tags associated with all the three images in the topic file.

Since the number of tags associated with each image is relatively small (6 on average according to our statistics), we think it would be helpful to extend the tags for the images to perform retrieval. We perform the tag expansion similarly as in section 2 and the results are listed in Tab. 8, where runs with the prefix WE denote the Wiki expanded version of the corresponding runs.

Fig. 3 illustrates the comparision of the results in Tab. 7 with that in Tab. 8. We can see that tag expansion really improve the performance greatly (with only one exception that for Image(all_3), which decrease about 3 percent on MnAP matric). We explain this as follows: the tags for the images may not be so accurate as that in the title field. So the expanded terms may drift away seriously. But if the title terms is also present, the expanding precess will be directed by them and the performance will increase finally as shown by the results of Title&Image(all_3) V.S. WE_Title&Image(all_3).

From Tab. 7 and Tab. 8, we can see that using the title filed as query perform better than using the tags of the images in the image fields, both for the original tags and for the Wiki Expanded ones. This means that the title field in the topic has more strong descriptive ability than the tags of the images. We can also see that when combining the title part with the image part to form the queries, the retrieval performance can always be greatly improved. This phenomenon is in

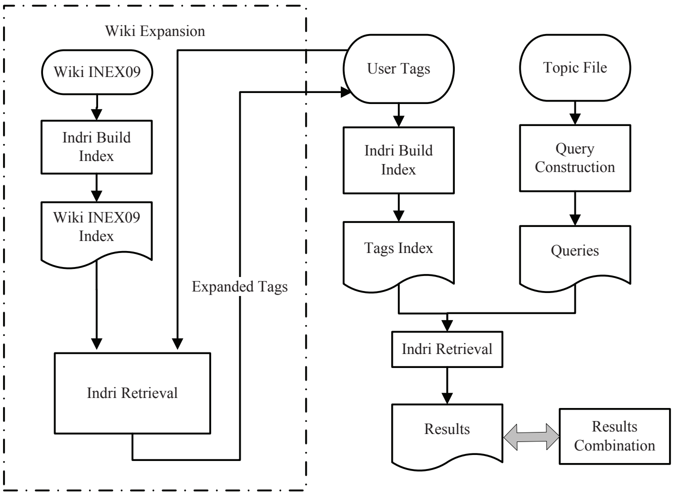


Fig. 2. Flowchart of the Tag-based Retrieval System

Table 7. Concept Retrieval with User Tags as Features(Without Wiki Expansion)

Result_name	**MnAP**	**AP@10**	**AP@20**	**AP@100**
Title	0.0802	0.0136	0.0376	0.1651
Image(all_3)	0.0763	0.0123	0.0320	0.1439
Image(first_not_null)	0.0711	0.0135	0.0241	0.1255
Title&Image(all_3)	0.0852	0.0137	0.0262	0.1635

accordance with our intuition that the more information provided, the better we can determine the user's Information Need.

We also have performed experiment to test the combination method at result level, which is denoted by WE_Combine_Title&Image(all_3)_at_Result_Level. To do this, we first generate the ranked Image Lists (WE_Title and WE_Image(all_3)) by the two methods independently. After normalizing the score for each Image *Im* to the same scale $(0-1)$, we re-rank the documents according to equation (7) [13]:

$$score_{\mathrm{T+I}}(Im) = w_{\mathrm{T}} * norm\_score_{\mathrm{T}}(Im) + w_{\mathrm{I}} * norm\_score_{\mathrm{I}}(Im) \ , \qquad (7)$$

where w_{T} and w_{I} are the combination parameters for Title and Image field whose ratio is $1:1$ for simplicity in our experiment. More reasonable ratio can be tested in the future.

Comparing the result to that at query level (WE_Title&Image(all_3)), we can see that they achieve similar performance.

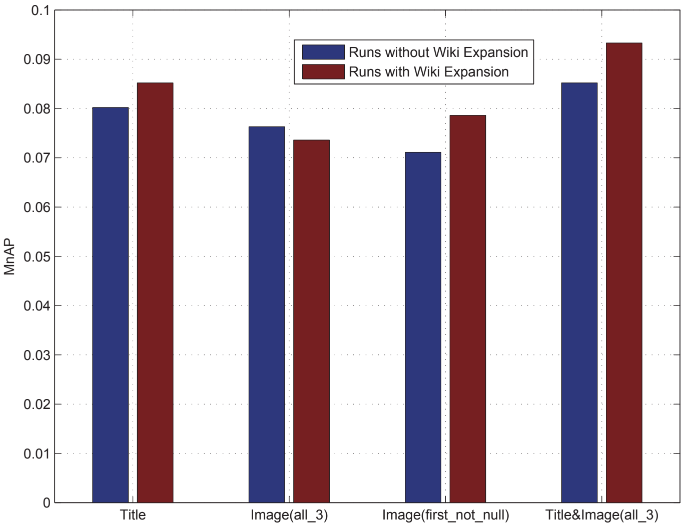


Fig. 3. Tag-based Retrieval with and without Wiki Expansion

3.2 Visual Feature based Retrieval Method

Extraction of Visual Features For visual feature extraction, we take the same method as that for visual concept annotation. So we omit their description here to reduce the lenght of the paper. Please refer to section 2 for more detail.

Classification We use the weighted distance from the query images in the topic file to the image being considered as the ranking score, which can be shown by (8):

$$RankingScore_f(Q, T_i) = \sum_{j=1}^{n} w_j dist_f(T_i, Q_j) \quad , \tag{8}$$

where Q denotes the query and T_i denotes the test image being considered; Q_j means the j^{th} image that belongs to the <image> field in the queries.xml file for query Q and n is the totoal number of <image> fields for query Q.

Tab. 9 lists the result of our visual based retrieval submission, which shows that by now our method does not achieve good performance either.

4 Conclusions

In this paper, we described the experiments we have performed for the "Visual concept detection, annotation, and retrieval using Flickr photos" task in detail.

Table 8. Concept Retrieval with User Tags as Features(With Wiki Expansion)

Result_name	**MnAP**	**AP@10**	**AP@20**	**AP@100**
WE_Title	0.0852	0.0187	0.0383	0.1721
WE_Image(all_3)	0.0736	0.0119	0.0212	0.1414
WE_Image(first_not_null)	0.0786	0.0133	0.0260	0.1311
WE_Title&Image(all_3)	0.0933	0.0187	0.0338	0.1715
WE_Combine_Title&Image(all_3) _at_Result_Level	0.0799	0.0141	0.0372	0.1638

Table 9. Concept Retrieval with Bag-of-Visual Words as Features

Result_name	**MnAP**	**AP@10**	**AP@20**	**AP@100**
BoV_Retrieval	0.0045	0.0030	0.0064	0.0316

We based our methods mainly on the Tags annotated by the users since we believe that there is strong relationship between the user's tag and the presence of a concept for annotation and between the user's tag and the query for retrieval. Official evaluation show that we achieved satisfied results for the User Tag based methods. Beside the main submission, we also perform some initial visual feature based experiments. However, the results we can achieved by now are not very good.

5 Future Works

Since this is the first time we participated the ImageCLEF task, we just did some initial work. More detailed experiments should be performed in the future. For example, for the Concept Annotation subtask, we didn't consider the relationships between different concepts. In reality, there are correlations between different concepts, such as the probability of the presence of timeofday_day is usually low given the presence of timeofday_night whereas the probability of view_outdoor will be high given the presence of the concept flora_tree. For the Concept-based Retrieval subtask, we just applied the traditional Language Modeling approach to the tag modeling application but did not take the specific characteristic of image tags into consideration, like that the terms in the documents in text retrieval are sufficient in general whereas the amount of tags for each image are relatively small, even after being expanded. More refined modification should be made to address these problems. We will do all these works in the future.

Acknowledgments. We would like to thank Prof. Jian-Yun Nie of Université de Montréal for his invaluable suggestions.

References

1. S. Nowak, S., K. Nagel, K., and J. Liebetrau. J.: The CLEF 2012 Photo Annotation and Concept-based Retrieval Tasks. In: CLEF 2012 Working Notes. Rome (2012)
2. Carpineto, C. and Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys(CSUR), 44(1), pp. 1-50 (2012)
3. Schenkel, R., Fabian M. Suchanek, and Gjergji Kasneci: YAWN: A Semantically Annotated Wikipedia XML Corpus. In: Proceedings of Datenbanksysteme in Business, Technologie and Web (BTW), pp. 277-291. Aachen (2007)
4. Metzler, D., Strohman, T., Zhou, Y., Croft, W.B.: Indri at TREC 2005: Terabyte Track. In: 14th Text Retrieval Conference, pp. 175-180. Gaithersburg (2005)
5. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Informational Retrieval, 11(2), pp. 77-107 (2008)
6. Chatzichristofis, S. and Boutalis, Y.: FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval. In: Proceedings of the Ninth International Workshop on Image Analysis for Multimedia Interactive Services, pp. 191-196. Klagenfurt (2008)
7. Mathias, L., Chatzichristofis, S.: Lire: Lucene Image Retrieval An Extensible Java CBIR Library. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 1085-1088. Vancouver, Canada (2008)
8. Yavlinsky, A., Pickering, M., Heesch, D.,Rüger, S.: A comparative study of evidence combination strategies. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 17-21. Montreal, Canada (2004)
9. Manning, C.D., Raghavan, P., and Schütze, H. Introduction to Information Retrieval. Cambridge University Press, Cambridge, UK (2008)
10. Metzler, D. and Croft, W.B.: Combining the Language Model and Inference Network Approaches to Retrieval. Information Processing and Management, 40(5), pp. 735-750 (2004)
11. Zhai, C.X. and Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS), 22(2), pp. 179-214 (2004)
12. Callan, J.P., Croft, W.B. and Harding, S.M.: The inquery retrieval system. In: Proceedings of the Third International Conference on Database and Expert Systems Applications, pp. 78-83, Spain (1992)
13. Croft, W.B.: Combining Approaches to Information Retrieval. Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Croft, W.B. (eds.), pp. 1-36. Kluwer Academic Publishers (2000)