Landmark Image Annotation Using Textual and Geolocation Metadata

Mădălina Mitran
IRIT, Paul Sabatier University
Toulouse, France
Madalina.Mitran@irit.fr

Rada Mihalcea
University of North Texas
Denton, TX, U.S.A
rada@cs.unt.edu

Guillaume Cabanac
IRIT, Paul Sabatier University
Toulouse, France
Guillaume.Cabanac@irit.fr

Mohand Boughanem
IRIT, Paul Sabatier University
Toulouse, France
Mohand.Boughanem@irit.fr

ABSTRACT

In this paper, we address the problem of landmark image annotation, defined as the task of automatically annotating a landmark query image with relevant descriptors (keywords or tags). Given a new query image along with its geolocation metadata (latitude and longitude), we retrieve several other images already available in a community image database (e.g., `flickr.com`, `panoramio.com`), found within a fixed radius of the location of the query image. We then formulate the automatic landmark image annotation problem as a tag ranking problem over all the tags obtained from these pre-existing neighboring images. We propose several tag ranking factors, and by evaluating them against a gold standard constructed using the geolocation-oriented photo sharing platform panoramio.com, we show that an aggregated measure that combines both distance and frequency factors leads to results significantly better than any of the individual factors.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: On-line Information Services

Keywords

landmark image annotations, social multimedia, collective knowledge, tag ranking, rank aggregation, Panoramio

1. INTRODUCTION

The development of digital technologies (e.g., digital cameras, smartphones) and the evolution of pervasive computing platforms have led to an increase in the number of images stored on our computers and on the web. This large amount of available images raises questions concerning their classification, exploitation, annotation, and retrieval. It is already established that users prefer to use text queries in order to retrieve documents online or elsewhere, and thus in order to retrieve digital images we should rely on the same process to which users are accustomed. The typical method used to access online images is to describe them using textual representations. With this aim, four approaches are presented in the literature: *a*) content-based approaches; *b*) manual annotation; *c*) semi-automatic annotation; and *d*) automatic annotation approaches. Several limitations hamper these approaches, including the inability of a machine to fully understand and interpret an image based on the low level visual features (also known as the "semantic gap" [12]), the lack of specificity, and the high cost of manual annotations. These issues are often addressed by methods that attempt to assist users during the image annotation stage, by developing automatic and semi-automatic photo annotation systems [8].

In this paper we formulate the automatic image annotation problem as a tag ranking probleme and we propose several tag ranking factors based on tag frequency, inverse tag frequency, and distance, and consequently measure the quality of the candidate tags obtained for a landmark query image. Our work exploits both the users social contributions (e.g., tags) as well as the metadata stored in the EXIF format [7] (e.g., latitude, longitude). To access this information we rely on photo sharing platforms such as `flickr.com` and `panoramio.com`, which gained huge popularity in recent years. For instance, Flickr (`flickr.com`) reached more than 51 million registered members[1].

In order to evaluate our work we created a dataset and a gold standard for a set of 30 query landmark images[2] (available upon request) relying on the photo sharing platform `panoramio.com`. Experiments performed against this dataset show that an aggregated measure that combines both distance and frequency factors leads to results significantly better than any of the individual factors.

The paper is structured as follows. Section 2 reviews related work, followed by Section 3, which describes several factors we propose to rank image tags. Section 4 presents the details of our evaluation, including experimental settings and results. Finally, Section 5 concludes the paper and proposes ideas for further work.

[1] http://advertising.yahoo.com/article/flickr.html

[2] A map with the 30 landmark images can be found at the following address: http://goo.gl/maps/ULrZB

2. RELATED WORK

To describe an image with relevant annotations, previous work was based on various features such as: visual features, textual features [1], user's contextual features [8], spatial features, and temporal features [11, 6].

Several previous works described the use of such features in the photo tag recommendation process together with two-three initial tags assigned by the users [10, 9]. Kucuktunc et al. proposed an automatic photo tag expansion system, using visual and textual features from other related images [9]. On the other hand, Sigurbjörnsson and van Zwol used tag co-occurrence statistics in order to recommend annotations for partially tagged photos in [10]. While this previous research made recommendations based on the few initial tags added by the users, in our approach we chose not to involve the user in the selection process, in order to ensure the scalability of the method, and also avoid any irrelevant initial tags that can lead to bad results.

In a related line of work, instead of asking the users for the initial tags, researchers explored the idea of using the title of an image together with the content similarity to retrieve related images together with their tags [1]. Other authors proposed an automatic approach that exploits the semantic correlation between image content and tags using Kernel Canonical Correlation Analysis [14].

The works that are perhaps most closely related to ours is that of Silva and Martins [11] and Moxley et al. [5]. Silva and Martins presented methods for annotating georeferenced photos with descriptive annotations available in online repositories such as `flickr.com` [11]. Their approach uses a set of estimators similar to our factors, but they did not make the difference between a general landmark annotation (which occurs in most images from a large area) and a specific landmark annotation. For example, for an Eiffel Tower photo, the general annotations can be "France" or "Paris," while annotations like "Tower" and "Eiffel Tower" represent specific annotations.

In contrast, Moxley et al. presented a tag suggestion tool, SpiritTagger, to suggest tags that reveal an insight into the spirit of a city or region [5]. They reranked tags considering their local frequency in comparison to their global frequency. Our general annotation differs from their global annotation in that our general annotation is represented by a small area neighboring the landmark represented by the query image and not by all the annotations available in a large area, such as Los Angeles and Southern California, as considered in [5]. Furthermore, we capture the specificity of a tag by using the inverse tag frequency, similar with the inverse document frequency (*idf*) used in Information Retrieval, and we rerank tags in order to boost the tags that are specific to a landmark. This aspect is not covered in their work and neither in that of Silva and Martins [11].

3. TAG RANKING FACTORS

As in [14], we address the problem of landmark image annotation as a tag ranking problem. With this aim, in this section we describe the factors used to rank tags.

We use the following notations in this section: *a)* p_q represents a landmark query image; *b)* p_r represents a retrieved photo $p_r \in P_r$; *c)* P_r represents the set of images retrieved for a landmark query image ; *d)* $T_{p_r} \in T_{p_q}$ represents the set of tags for a retrieved photo p_r; *e)* T_{p_q} represents the set of tags for a query image derived from the retrieved photos; and *f)* t represents a tag $\in T_{p_q}$ to which we apply different factors.

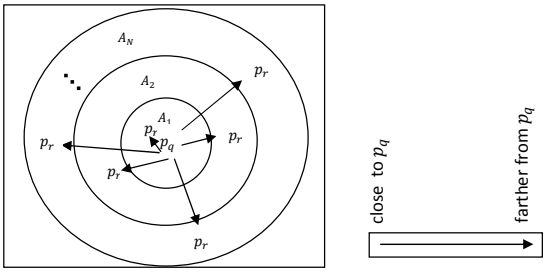

Figure 1: Areas division according to a fixed radius.

3.1 Tag Frequency Factor

The tag frequency factor represents the number of times a tag appears in the result list (T_{p_q}) of a landmark query image p_q. This factor is similar with the tf measure from Information Retrieval and it is calculated according to the following equation:

$$m_1(p_q, t) = tf(t, T_{p_q}) \tag{1}$$

3.2 Inverse Tag Frequency Factor

We use the inverse tag frequency factor (similar to the *idf* from Information Retrieval) to determine the importance of a candidate tag according the level of specificity across the entire corpus. We propose two methods to calculate the inverse tag frequency factor (i.e., internal and external).

3.2.1 Internal Inverse Tag Frequency Factor

In order to calculate the internal inverse tag factor we use an internal dataset (see Section 4 for the dataset used) and we divide the query image area in several areas: A_1, A_2, $\cdots$, A_N (see Figure 1). This factor is calculated according to the following equation:

$$m_2(p_q, t) = idf_{intern} = log\frac{N + 0.1}{n} \tag{2}$$

where N represents the number of areas resulting from the division of the query image area and n represents the number of areas where a tag t occurs. We added 0.1 to the numerator to avoid zeros for tags that occur in all the areas.

3.2.2 External Inverse Tag Frequency Factor

The external inverse tag frequency factor is calculated by using an external corpus. We use the "keyphraseness" method presented by Mihalcea et al. in [4] as an estimate for the External Inverse Tag Frequency Factor, which is estimated based on counts obtained from Wikipedia:

$$m_3(p_q, t) = idf_{extern} = Keyphraseness \tag{3}$$

3.3 Distance Factor

The aim of this factor is to capture the importance of a candidate tag. We make the assumption that a tag is more important for a landmark query image if it occurs in images located in areas close to the location of the landmark represented by the query image (i.e., the distance between a query image and a retrieved image is small). For example, in Figure 1, tags that occur in areas close to p_q, such as A_1

and A_2, are more important than tags that occur in distant areas like A_N. In the following, we propose two methods to represent the distance factor (i.e., physical distance and id area distance).

3.3.1 Physical Distance Factor

The physical distance factor is represented by the distance (in meters) between a query image and a retrieved image. We compute a geospatial similarity gs between a query image and each retrieved image from P_r as in equation 4.

$$gs(p_q, p_r) = \frac{1}{d(p_q, p_r)}, \quad (4)$$

where d represents the distance in meters between two geographical points in terms of latitude and longitude. To compute this distance we use the great circle method[3], used also in [11]. As each retrieved photo p_r is represented by a set of tags T_{p_r}, each tag from this set receives a score represented by the geospatial similarity aforementioned. Therefore, the final score for a tag t is computed as in equation 5.

$$m_4(p_q, t) = \sum_{p_r \in T_{p_r}} gs(p_q, p_r) \quad (5)$$

3.3.2 Id Area Distance Factor

This distance is represented by the id of an area. For example, for a tag t that occurs in the A_1 and A_3 areas, the distance will be represented as either 1 or 3 (see Figure 1). Therefore, the final score for a tag t is computed as in equation 6.

$$m_5(p_q, t) = \sum_{i=1}^{q} \frac{tf_q}{idArea_q}, \quad (6)$$

where q represents the number of areas obtained from the division of the query image area according to a fixed radius, tf_q represents the number of times a tag occurs in an area q, and $idArea_q$ represent the id of an area.

4. EVALUATION

In this section, we evaluate how well the different tag ranking factors, as well as their combinations, can describe a landmark query image.

4.1 Experiments Setup

Dataset. We built a dataset of 30 landmark query images from all over the world. The 30 query images were chosen to ensure that a sufficient number of images can be found in their neighboring. We chose to create our own dataset mainly because the other datasets publicly available do not model the contextual information used in our work (for example, the Corel dataset does not contain any contextual data). Using the photo sharing website `panoramio.com`, for a query image p_q and a radius r (in meters), we retrieve the images found in the area having as origin the point of latitude and longitude of p_q and a radius of r. The retrieved images together with their metadata (image id, image title, image tags, latitude, longitude, date/time, and the image url) are stored in an Oracle database. By removing noisy tags that are not suitable for our propose (e.g., numeric tags, empty tags, special characters), we obtain a dataset containing 40,366 distinct images and 6,144 unique tags. The average number of distinct tags per image is 6.56.

[3]http://en.wikipedia.org/wiki/Great-circle_distance

Table 1: Evaluation results for each individual factor

Criteria	P@3	P@5	P@10	MAP	MRR
m_1	**0.788**	**0.706**	**0.523**	**0.490**	**0.983**
m_2	0.033	0.026	0.023	0.032	0.092
m_3	0.220	0.220	0.203	0.169	0.404
m_4	0.655	0.626	0.490	0.440	0.784
m_5	0.611	0.600	0.483	0.427	0.789

Gold Standard. The gold standard was created with the help of 32 researchers, all of them PhD or masters students in our computer science department. The evaluators had to accomplish two tasks: *a*) First, choose six images they knew best from the 30 landmark query images. We did not randomize this process because we wanted to be sure that each assessor knows well the landmarks represented in the query images in order to make good judgments. *b*) Second, for each selected image, they had to select only the tags that were good image descriptors.

After collecting the gold standard annotations, we used the CombMNZ method [2] to combine the individual assessor ranks obtained for a query image into a single rank, where a tag is even more relevant if it is identified as such by a large number of assessors. The relevance of this method was shown by Lee [3] on TREC lists combination results.

Metrics. The performance is measured by three metrics traditionally used in Information Retrieval: *a*) Precision at rank k (P@k). We calculate the precision at rank 3, 5, and 10 (P@3, P@5, P@10); *b*) Mean Average Precision (MAP); and *c*) Mean Reciprocal Rank (MRR).

4.2 Results and Discussion

4.2.1 Performance of individual factor

Table 1 presents the results for each factor using the gold standard and the dataset obtained from `panoramio.com`. The results show that the tag frequency factor has the best results with a precision P@5 of 70%. This result was expected because the most frequent tags from a dataset have the tendency to be relevant and usually they represent the general tags for a landmark query image. We also notice (Table 1) that both distance factors perform well, with a precision P@5 of 62% and 60% respectively. Therefore, we expect to improve our results by aggregating these factors.

4.2.2 Performance of aggregation factors

In this section we present eight aggregation methods together with their results.

- $AggM1$ based on the tag frequency factor and the internal inverse tag frequency factor. This method is similar to the $tf * idf$ measure in Information Retrieval and it is calculated according to the equation: $AggM2 = m_1 * m_2$;
- $AggM2$ based on the tag frequency factor and the external inverse tag frequency factor (similar to $AggM1$): $AggM2 = m_1 * m_3$;
- $AggM3$ based on the tag frequency factor, the internal

Table 2: Evaluation results for the eight aggregation methods (AggM).

AggMeth	P@3	P@5	P@10	MAP	MRR
AggM1	0.522	0.493	0.403	0.341	0.716
AggM2	0.855	0.740	0.543	0.496	**1.000**
AggM3	0.444	0.380	0.353	0.325	0.662
AggM4	0.522	0.466	0.396	0.362	0.771
AggM5	0.688	0.646	0.513	0.444	0.819
AggM6	0.888	**0.773**	**0.573**	**0.541**	**1.000**
AggM7	0.644	0.533	0.430	0.389	0.851
AggM8	**0.889**	0.746	0.566	0.515	0.983

inverse tag frequency factor, and the physical distance in meters: $AggM3 = m_2 * m_4$;

- $AggM4$ based on the tag frequency factor, the internal inverse tag frequency factor, and the distance represented by the id of an area: $AggM4 = m_2 * m_5$;
- $AggM5$ based on the tag frequency factor, the external inverse tag frequency factor, and the physical distance in meters: $AggM5 = m_3 * m_4$;
- $AggM6$ based on the tag frequency factor, the external inverse tag frequency factor, and the distance represented by the id of an area: $AggM6 = m_3 * m_5$;
- $AggM7$ based on the tag frequency factor, the internal inverse tag frequency factor, the id of an area, and the number of areas obtained from the division of a query image area: $AggM7 = m_2 * \sum_{i=1}^{q} m_{1q} * (nbArea - rank_q + 1)$, where m_{1q} represents the number of times the tag t appears in the q area, $nbArea$ represents the number of areas, and $rank_q$ represents the id of the current area;
- $AggM8$ similar to the $AggM7$ method, the difference being that for this aggregation we use the external inverse tag frequency factor: $AggM8 = m_3 * \sum_{i=1}^{q} m_{1q} * (nbArea - rank_q + 1)$.

When we combine both distance and frequency factors, we obtain results significantly better than any of the individual factors, as shown in Table 2. For instance, regarding best P@5 among all factors (i.e., $m_1 = 0.706$) versus all aggregation methods (i.e., $AggM6 = 0.773$), we found a 9.5% increase in effectiveness, which is statistically significant according to Student's bilateral and paired t-test [13] with $p < 0.05$. Therefore, the best performance is obtained for the aggregation of the following three factors: tag frequency, external inverse tag frequency, and area distance. As the precision P@5 of this aggregation method goes up to 77%, it means that on average 3.85 of the top 5 tags are good descriptors for a query image. We also notice (Table 2) that the external inverse tag frequency factor performs better that the internal inverse tag frequency factor for all the aggregation methods in terms of identifying the top N tags for a query image (P@3, P@5, P@10).

5. CONCLUSIONS

In this paper, we addressed the problem of landmark image annotation, which we formulated as a tag ranking problem. The set of tags available for a landmark query image are ranked according to several ranking factors. Experimental results based on a real dataset obtained from a photo sharing platform show that an aggregation measure that combines both distance and frequency factors performs better than any of the individual factors.

Future work will include studies on other available resources for landmarks images. A possible resource can be the `wikipedia.org` platform: by using the location metadata, we can identify the article that describes the landmark query image and use that information to improve our tag ranking factors. Furthermore, a synonym database can be used to expand the tags of a query image. All these additional resources may better reflect the relevant descriptors for a landmark query image. Finally, we would like to also experiment on larger test and training datasets.

6. REFERENCES

[1] S. Barai and A. F. Cardenas. Image annotation system using visual and textual features. In *DMS '10*, pages 289–296, 2010.

[2] E. A. Fox and J. A. Shaw. Combination of multiple searches. In *TREC-1: Proceedings of the First Text REtrieval Conference*, pages 243–252, 1993.

[3] J. H. Lee. Analyses of multiple evidence combination. In *SIGIR '97*, pages 267–276, New York, USA, 1997.

[4] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, pages 233–242, NY, USA, 2007.

[5] E. Moxley, J. Kleban, and B. S. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *MIR '08*, pages 24–30, NY, USA, 2008.

[6] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *MULTIMEDIA '04*, pages 196–203, NY, USA, 2004.

[7] T. S. C. on AV & IT Storage Systems and Equipment. Exchangeable image file format for digital still cameras: Exif version 2.3. Technical Report JEITA CP-3451B, April 2010.

[8] A. Rae, B. Sigurbjörnsson, and R. van Zwol. Improving tag recommendation using social networks. In *RIAO '10*, pages 92–99, 2010.

[9] S. G. Sevil, O. Kucuktunc, P. Duygulu, and F. Can. Automatic tag expansion using visual similarity for photo sharing websites. *Multimedia Tools Appl.*, 49(1):81–99, 2010.

[10] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08*, pages 327–336, 2008.

[11] A. Silva and B. Martins. Tag recommendation for georeferenced photos. In *LBSN '11*, pages 13:1–13:8, 2011.

[12] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:1349–1380, 2000.

[13] Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908.

[14] Z. Wang and B. Li. Learning to recommend tags for on-line photos. In *2nd International Workshop on Social Computing, Behavior Modeling, and Prediction*, 2009.