Quest for Relevant Tags using Local Interaction Networks and Visual Content

Neela Sawant, Ritendra Datta[*], Jia Li, James Z. Wang
The Pennsylvania State University, University Park, Pennsylvania, USA
{neela, datta, jiali, jwang}@psu.edu

ABSTRACT

Typical tag recommendation systems for photos shared on social networks such as Flickr, use visual content analysis, collaborative filtering or personalization strategies to produce annotations. However, the dependence on manual intervention and the knowledge of sufficient personal preferences coupled with the folksonomic issues limit the scope of these strategies. In this paper, we present a fully automatic and folksonomically scalable tag recommendation model that can recommend tags for a user's photos without an explicit knowledge of the user's personal tagging preferences. The model is learned using the collective tagging behavior of other users in the user's local interaction network, which we believe approximates the user's preferences, at least partially. The tag recommendation model generates content-based annotations and then uses a Naïve Bayes formulation to translate these annotations to a set of folksonomic tags selected from the tags used by the users in the local interaction network. Quantitative and qualitative comparisons with 890 Flickr networks show that this approach is highly useful for tag recommendation in the presence of insufficient information of a user's own preferences.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data sharing*

General Terms

Algorithms, Experimentation, Human Factors

Keywords

image annotation, tag recommendation, social media, local interaction networks, user preference approximation, folksonomy, Flickr, ALIPR

1. INTRODUCTION

Uploading and sharing photos with friends and the world has become a common practice, facilitated greatly by the online photo-sharing tools and the social networking platforms such as Yahoo! Flickr [28], Google Picasa [9], Photobucket [22], Facebook [4], and Orkut [21]. Facebook, for example, hosts over ten billion photos uploaded since its launch [5], with continued rapid growth in volume. Flickr also serves over three billion images [6]. The users not only upload photos, but also provide tags, which act as annotations to the photos, making them easier for retrieval. This form of photo content distribution in social networks is becoming one of the major drivers of current image retrieval infrastructures. Yahoo! [29], for example, integrates Flickr images in its keyword based web image search.

The success of tag based retrieval depends on the individual choices to annotate images, and therefore, is affected by the personal tendencies and community influences [23]. Folksonomic issues of polysemy, synonymy and basic level variation [8], in addition to the occasional disparity between the annotations and the visual content [26] impact the relevance of tags to images. Further, a recent study [25] shows that the number of tags, and moreover that of useful tags, is low. To deal with these challenges, in recent years multimedia researchers have been concentrating on automatically annotating images or recommending tags using techniques such as visual content analysis, user personalization modeling, collaborative filtering, or a combination thereof.

Pure visual content based methods involve learning the association between visual features and words to automatically generate textual annotations for images. Though the relevance of such results may be high, the training efforts involved in manually creating high quality training sets limit these methods to a relatively small number of visual concepts and do not scale to the millions of folksonomic tags. Collaborative filtering based methods, on the other hand, can handle a very large folksonomic tag set, but typically require the users to specify at least one tag per image initially. Clearly, for completely untagged images, this strategy cannot produce annotations. Furthermore, when related tags are computed using a global co-occurrence, interesting and locally contextual tags may be suppressed. Personalization based methods that model the user's personal tagging preferences by observing the tagging history, can produce locally relevant tags. However, for users without sufficient tagging history or having many idiosyncrasies, personalization may not be very helpful. (We discuss some of these systems in Section 2.) In our opinion, due to these limiting factors, a

[*] Ritendra Datta is currently affiliated with Google Inc., Pittsburgh, Pennsylvania, USA.

MIR'10, March 29–31, 2010, Philadelphia, Pennsylvania, USA.
Copyright 2010 ACM 978-1-60558-815-5/10/03 ...$10.00.

large number of untagged images still lie in obscurity and alternate strategies need to be devised to produce visually relevant and contextual tags.

In this paper, we propose an image tag recommendation strategy that is fully automatic, incorporates visual content and yet scales to the large folksonomic vocabularies. The main research problem addressed is that of an insufficient knowledge of a user's personal tagging preferences, which is approximated by modeling the tagging behavior of other users with whom the user has interacted. We refer to this set of other users as the *local interaction network* (LIN) and the user under consideration is referred to as the *seed* of the local interaction network. This idea is inspired from the idiom that *'birds of a feather flock together'*, i.e., the belief that the users belonging to a social network may share similar interests, vocabulary and behavior. Therefore, it may be possible to characterize a user, at least partially, by collectively analyzing the photo tagging behavior of that of her social network.

An obvious application can be envisioned in the case of special interest groups on the photo sharing websites such as Flickr. Flickr groups are self-organized groups of users, typically with common interests and themes. Different types of groups can form among geographically co-located users (e.g., university students) or users with particular topic interests (e.g., dogs) or photographic interests (e.g., macro, still). If a user belongs to the special interest group, say *'I love dogs'*, it may be reasonable to hypothesize that the visual content of the user's photos may be annotated using tags such as 'dog', 'puppy', 'Labrador', without looking at the actual tags given by the user. Thus, the characterization of the social network provides a context for predicting a user's tagging behavior. Similar principle can be observed in the topic-based representation of Flickr groups [20] which shows that coherent themes can be identified by analyzing the collective behavior of users in the same group.

However, the concept of a user being similar to her network does not need to be restricted to explicitly defined groups. In this paper, we validate our hypothesis by using simple interactions among users such as commenting on each other's photos, to form interaction-based networks and show that these can be quite useful in approximating a user's tagging behavior. The collective tag vocabulary of the users with whom a user interacts, incorporates a significant portion of the user's own tag vocabulary. We further compare different strategies to select a practical number of tags from the collective tag vocabulary of the network and train a Bayesian model to combine them with the predictions of a content based image annotation system, ALIPR (Automatic Linguistic Indexing of Pictures - Real time [13]). This way, the limited number of visual content based annotations translate into arbitrarily large folksonomies. Experiments on a large number of Flickr networks and comparisons with other methods show the applicability of this technique. We conclude that even when a user's personal preferences are not discernible, an approximation can be made from the tagging behavior of the user's local interaction network and relevant tags can be suggested in a fully automated manner. This real world problem is highly complex owing to the myriad factors that characterize social networks and associate meanings to the actions of the people. Statistical measures and formulations are instrumental for sense-making in such scenarios and facilitate the identification of the relevant information even in the presence of noisy tags and activities.

The paper is organized as follows: Section 2 presents the related work in image tag recommendation. Section 3 introduces the details of the local interaction networks, that are further explored in Section 4, to select a set of useful tags for the approximation of a user's own tags. Section 5 details the Naïve Bayes framework for combining the content based image annotations with the selected set of network tags. Section 6 shows the experimental results, both quantitative and qualitative. Discussion and conclusions are presented in Section 7 and Section 8 respectively.

2. RELATED WORK

Automated image annotation has been a topic of interest since the early 2000s. In this section, we provide some references for the image tagging techniques developed over the years. A comprehensive list of references on content based image retrieval and annotation can be found in [3] and [26].

Before the advent of the Web 2.0, most automatic photo annotation systems relied on pure visual content analysis. Systems such as [1, 10, 12, 13] use statistical modeling techniques to identify the visual concepts in images. Bernard et al. [1] propose an annotation scheme using multi-modal data mining over image regions and text. Jeon et al. present a 'cross-media relevance model' [10] which directly computes the probability of annotations using the blob-based visual features of an image. *ALIPR* [13] is a recent system for real-time annotation of images based on the probabilistic modeling of color and texture features from a set of visual categories. The controlled vocabulary of visual concepts, used in these and many other content based annotation methods is insufficient to handle the large number of real-world folksonomic tags.

Two systems that can potentially handle unlimited vocabulary are presented in [14] and [15]. Li et al. present a content based image retrieval system [14] that propagates textual features from visually similar images. Lindstaedt et al. [15] use a combination of off-line supervised classification of images into selected concepts followed by tag propagation from visual similar images. However, these systems need to analyze very large datasets and do not address the localized relevance of tags in a folksonomic setting.

The major body of current work in tag recommendation falls into the category of semi-automatic systems, due to the requirement on the user's part to manually provide an initial set of tags. Sigurbjörnsson et al. [25] and Garg et al. [7] present strategies to predict additional tags that have a frequent co-occurrence with the manually provided tags. The co-occurrence is computed over large user collections and may lose relevance in the local context of users. The *Tag Suggestr* [11] system based on a combination of visual and text features, uses the initial set of tags to retrieve additional tags from related photos that have some of the initial tags in their tag lists. The candidate tags are weighted using visual similarity between their original images and the user-uploaded photo. The *Annosearch* system [27] uses the initial set of tags to fetch images with similar textual descriptions. Annotations are then mined from the description of visually similar images. Yan et al. [30] propose a 'frequency based annotation' algorithm to speed up the manual annotation process by distinguishing between the more frequent tags for browsing and the less frequent (but more discriminative) tags for annotation. Shepitsen et al. [24] present a personalized recommendation system that incorporates user profiles and previous tag clusters to re-rank the tags suggested by a non-personalized recommendation algorithm.

In the context of social networks, a number of tagging tools have been devised to harness the rich content associated with photos. *SpiritTagger* [17], for example, utilizes the GPS coordinates of photos combined with the content based image analysis to annotate photos with other geographically relevant tags. *ZoneTag* [18] is another tool that uses GPS location to identify related tags from a user's social network. The *PeopleRank* algorithm [19] is specially designed to identify and annotate people in personal photos using time and location of photos. Lindstaedt et al. [16] present a *tagr* system based on a mash up of different image and user features.

2.1 Our Contribution

As aforementioned, a major contribution of this paper is in showing that simple interactions in social networking sites yield interaction based networks that can be used to build a useful tag recommendation framework. The process of tag recommendation is fully automatic and can extend the content based annotations to arbitrary folksonomic settings, even when the user's personal tag preferences are unknown.

The underlying belief in our approach that the users share similar meanings and context with their connections as opposed to the global user collection, is the basic difference between our approach and many existing approaches discussed earlier. We argue that the use of local networks for tag behavior approximation has the following advantages:

- The collective vocabulary of the local network is much less susceptible to personal idiosyncrasies. The application of a tag to an image may be arbitrary in case of one person, but such noisy tag applications can be suppressed if a systematic trend is observed in the application of tags to specific visual content over multiple users that are connected to the user.

- The context in which tags are applied or the interesting local semantics may get obscured in a global analysis comprising of many diverse users. However, local networks may be well able to capture the same.

3. LOCAL INTERACTION NETWORKS

In this section, we define the local interaction network (LIN) and introduce other terminology. Further, the test dataset used to evaluate the efficacy of the proposed approach is presented.

3.1 Definition

In general, a social network consists of a large number of user pair connections (u_i, u_j), whereby a user u_i and a user u_j are connected at some level. An explicit level of connectivity can be a pro-actively specified 'friendship' between u_i and u_j, applicable to the social networking sites like Facebook. Connections can also form due to the co-membership of u_i and u_j in some special interest groups.

Our definition creates a local comment based interaction network of users. A connection is created between users u_i and u_j when either u_i comments on the photos of u_j or vice versa. Let us denote the user whose photos are to be tagged as the seed user. Then, **for a seed user u_s, the local comment interaction network LIN is computed as the set of users $\chi(u_s) = \{u_i | u_i \neq u_s, u_i$ comments on u_s's photos or u_s comments on u_i's photos$\}$**.

The set of tags actively used by a user u_i to tag her own or others' photos is defined as the tag vocabulary T_{u_i} of that user. More explicitly, a tag t given by a user u_i on a photo belonging to another user u_j contributes to the vocabulary of only u_i and not u_j (unless of course u_j has herself used the same tag in tagging at least one photo).

If T_{u_i} represents the vocabulary for a user u_i (set of tags actively used by the user u_i), then the collective vocabulary $T_{\chi(u_s)}$ of a seed user u_s's network can be computed as:

$$T_{\chi(u_s)} = \bigcup_{\forall u_i \in \chi(u_s)} T_{u_i} \quad . \tag{1}$$

Theoretically the definition of local interaction networks gives rise to a complete comment-interaction graph of Flickr users. However, due to the limitations on the data collection, we may not be able to identify the complete set of connections unless all the photos from all users have been crawled (or a system keeps track of the users' online commenting behavior). For example, given a user u_i, we can analyze all of u_i's images to identify the set of other users $\{u_j | u_j \neq u_i; u_j$ comments on u_i's photos$\}$. However, given that same user u_i, we may not know the set of users $\{u_k | u_k \neq u_i; u_i$ comments on u_k's photos$\}$. Therefore, we work with the available interactions present in the crawled dataset and show that even with the partial interactions, users can be characterized. For practical considerations, we retain only the user pairs where one user comments at least three times (an empirical threshold) on the photos of another.

We acknowledge that a number of other interaction modes are facilitated by the social networks such as the photo or profile views (u_i views u_j's photos or profile), action of liking/ favoriting a photo (u_i likes u_j's photo or marks it as a favorite) or tagging collaboration (u_i tags u_j's photo or both of them tag a photo in common). However, connections formed through comments are more easily traceable (as opposed to photo or profile views) and quite abundantly available (as opposed to favoriting or tagging collaboration). Also note that it does not directly imply a similarity in the tags used by the connected users, but it merely hints at the possibility of shared interests.

3.2 Data Collection

The data collection strategy was driven by the definition of the local comment interaction networks. Starting with 1023 random user seeds, more users in their networks were identified from the comments on the seeds' photos. The photos for these additional users were crawled and comments were analyzed to incorporate any additional connections where previously unconnected two users from the existing user set were linked through comment interaction.

Fig. 1 shows the distribution of network sizes, i.e., of the number of users in a network, over the 1023 networks in the dataset. The minimum number of connections is two users with the median and the average at 16 and 42.497, respectively. This shows that all seed users had a comment interaction network, and therefore the approach is applicable in real world scenarios. In Section 6.3, we analyze if any discernible correlation exists between the network size and the efficacy of the proposed tag recommendation framework.

From the collected data, we identified the tag vocabulary

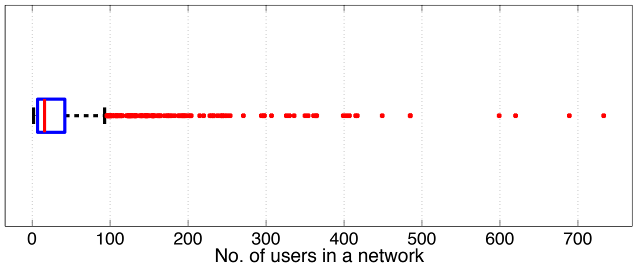


Figure 1: Boxplot for the network size, i.e., the number of users in the networks of the seed users.

for each user in the 1023 local interaction networks. Totally, 2,266,215 unique tags were collected of which 1,017,163 were used exactly once (44%). We removed these tags to simplify the analysis. Removal of these tags broke some network connections and rendered them degenerate. We chose to limit the experiment to the networks where the seed's vocabulary had at least 10 tags. In future, additional mechanisms such as stemming and stop-lists can be used to further reduce the folksonomic issues.

On the application of these simple filtering techniques, the revised data has the following attributes:

- 890 networks (corresponding to the 890 seed users)
- 27,708 total users
- ∼5.5M images
- ∼49.4M tag applications (tag-to-image assignments)
- ∼1.25M unique tags

It is evident that the problem is very complex and requires effective measures for useful image annotation.

4. EXPLORATORY ANALYSIS

In this section, we show that a seed user's interaction network indeed has the capability to emulate the seed's preferences, at least partially.

Fig. 2 shows the distribution for the percentage vocabulary set intersection between a seed's vocabulary and that of its network, i.e.,

$$\frac{|T_{u_s} \cap T_{\chi(u_s)}| \times 100}{|T_{u_s}|} \quad . \tag{2}$$

The mean vocabulary overlap over all seed users is 51%. This is significant, especially since there is no stipulation that a seed user's tags will be present in the corresponding network vocabulary. It indicates that the local interaction networks are significantly related to the seed users.

4.1 Analogy to Signal and Noise

Fig. 3 shows the distribution of the sizes of the network vocabularies. Comparing this with the total number of unique tags in the system (1.25M), it is clear that a relatively much smaller tag set needs to be searched to locate a significant fraction of tags relevant to the seed user's vocabulary.

Metaphorically, the relevant tags can be considered as the signal required for user characterization and the remaining unrelated tags as the noise present in the corresponding network vocabulary. If we define the signal to noise ratio (SNR) as

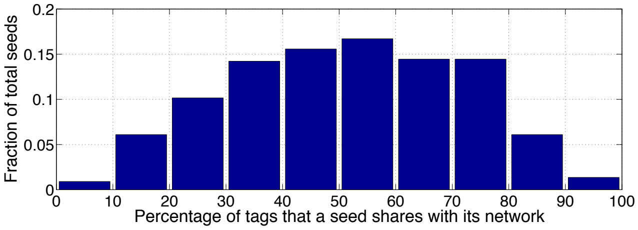


Figure 2: Distribution of the percentage common vocabulary of the seed and its network.

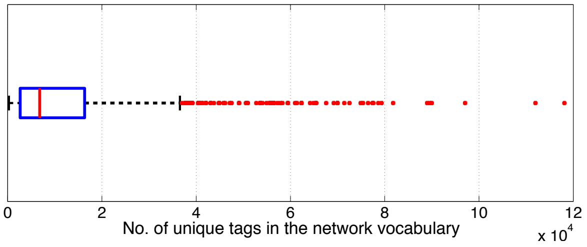


Figure 3: Boxplot for the size of network vocabulary.

$$SNR = \frac{|T_{\chi(u_s)} \cap T_{u_s}|}{|T_{\chi(u_s)}| - |T_{\chi(u_s)} \cap T_{u_s}|} \tag{3}$$

then statistically, the signal to noise ratio in the local interaction networks is significantly higher (p-value=0) as opposed to that computed using the global vocabulary comprising of all users in the system. Therefore, a user is much more similar to her own network and using the local network to characterize a user is a good idea in the absence of any knowledge of the user's personal preferences.

Still, as is evident from Fig. 3, some network vocabularies are very large to be of any practical use. Therefore, we would like to increase the SNR further, by removing additional irrelevant tags and selecting a practical subset of tags for each user under consideration.

4.2 Increasing SNR

Social networks are a complex phenomenon. The different characteristics of social networks may influence the extent to which the local interaction networks resemble the seed users. We explore four simple strategies with the purpose of selecting a practical set of tags from the network vocabulary corresponding to each of the individual seeds.

1. ALL (Complete network vocabulary) - All tags from the network vocabulary $T_{\chi(u_s)}$ (Eq. 1) are considered for experimentation. This is the baseline.

2. POP (Popular network tags) - Limiting the vocabulary to the most frequent words in $T_{\chi(u_s)}$ is a basic improvisation with the hope that the most frequent words of the network are likely to capture the corresponding seed user's interests. This strategy is likely to be sensitive to a bias from the more active users within the network who give a large number of tags as opposed to other users. For example, an idiosyncratic tag that

is used by one single user 10001 times will be ranked higher than a tag that is used by 1000 users in the network, ten times each.

For present experiments, we choose 3000 most frequently occurring tags. This number was chosen empirically as the range of the 890 seeds' vocabulary sizes ($|T_{u_s}|$) was found to be nearly 3000.

3. ATLST2 (Tags used by at least 2 users) - This strategy compensates for the idiosyncratic tags coming from single overly active users. The vocabulary is filtered to retain only those tags that are used by at least two users in the network. Therefore, a tag that is used 10000 times by one single user does not contribute to the network vocabulary, whereas a tag that is used once each by two different connections counts.

4. COMPOSITE (modified ATLST2+POP) - In this strategy, each word is weighted by the number of users who use it. The weighting factor is an arbitrarily large number that boosts the tags used by more than one user, but also preserves the frequency based order of tags within the set of tags used by the same number of users. After a composite ranking is thus obtained, only the top 3000 tags are retained. In this strategy, if the number of tags used by at least two users is less than 3000, tags that are used by one friend each, are pooled in the order of their respective frequency in the composite vocabulary. However, if the number of tags used by at least two users exceeds 3000, the tags that are used collectively by more number of friends get priority and tags that are used less frequently get cut off after the first 3000 tags have been collected.

For each of the four strategies, the corresponding network vocabulary was computed for each of the 890 seeds. To identify the overall superiority of a strategy, we compared the SNR of these four strategies across all seed users using paired t-tests. In a paired t-test, Strategy, say X, was compared with another strategy Y using the null hypothesis H_0**: SNR(X) = SNR(Y)** and the alternative hypothesis H_a**: SNR(X) < SNR(Y)** (or SNR(Y) > SNR(X)). In all combinations of the strategy pairs, the null hypothesis was rejected at the significance level of 0.05, to indicate a clear ordering of the strategies in terms of the ability of user approximation.

We repeated the paired t-tests for two separate categories of seed users: a) users whose ATLST2 network vocabulary had less than or equal to 3000 tags, and b) users whose ATLST2 network vocabulary had more than 3000 tags (and therefore, it had to be truncated in the COMPOSITE case).

Table 1 shows the summary of the paired t-test results that were useful in obtaining a ranking of strategies in the increasing order of SNR. From Table 1, it is clear that:

- In the first case, the ordering in terms of efficacy is: **ALL < POP < COMPOSITE < ATLST2**. Therefore, it is best to use the ATLST2 vocabulary whenever it has less than 3000 tags,

- A different order is obtained in the second case: **ALL < POP < ATLST2 < COMPOSITE**. Therefore, whenever the ATLST2 vocabulary has more than 3000 tags, it is best to truncate it by appropriating the COMPOSITE strategy.

This experiment makes it evident that for all practical purposes, we do not have to consider an arbitrarily large network vocabulary. In this case, up to 3000 tags are sufficient to get a useful characterization. We denote the selected network vocabulary using the above observations as $\tilde{T}_{\chi(u_s)}$.

5. RECOMMENDATION FRAMEWORK

In this section, we describe the tag recommendation framework that combines the predictions of a visual content based annotation system with the network vocabulary $\tilde{T}_{\chi(u_s)}$. We use annotations produced by ALIPR [13] which is a fully automatic and real-time image annotation system. To extend the limited ALIPR vocabulary to $\tilde{T}_{\chi(u_s)}$, we use the notion of inductive transfer or transfer learning, which was effectively used in the PLMFIT model in a recent work [2]. The idea is that the knowledge acquired using one set of training instances, (in this case, the tagged images used in ALIPR training) can be used to infer answers of other related questions. Our goal is to infer the user tags with the help of ALIPR generated tags, albeit indirectly. For example, if ALIPR very effectively learned to recognize the concept 'dog' from a function of visual features, but the photos of the network associate a related concept 'puppy' to that function, then the frequent co-occurrence of 'dog' in the ALIPR tag set and 'puppy' in the $\tilde{T}_{\chi(u_s)}$ indicates that an ALIPR prediction of a tag 'dog' should be translated to the tag 'puppy' for a better estimate of the actual user tags.

5.1 ALIPR

ALIPR uses a generative modeling technique to build a probability measure over the discrete distribution representation (i.e., bags of weighted vectors) of color and texture features. It is trained to detect 332 concepts such as manmade, art, sky, water, modern, flower, to name a few. Given an image, ALIPR produces an ordered vector of possible annotations (and corresponding probabilities) in real time. For the tag recommendation framework, the ALIPR annotations for each image are converted to a binary vector of 332 dimensions $(a_1, ..., a_N)$, $N = 332$. If the probability $P(a_i)$ of the i^{th} ALIPR concept is greater than a reasonable threshold (such as 0.1), a_i is set to 1, otherwise $a_i = 0$.

5.2 Extending ALIPR Annotations

Let u_s represent the seed user under consideration and $\chi(u_s)$ its local interaction network. Let $\mathcal{I}_{\chi(u_s)}$ be the set of the training images belonging to all users in $\chi(u_s)$. If $\tilde{T}_{\chi(u_s)} = \{t_i | i = 1, \ldots, M\}$ denotes the network vocabulary over which the network's tag distribution $P^{\chi}(\cdot)$ is computed, then, the probability of observing a network tag t_i, $P^{\chi}(t_i = 1)$ can be estimated as,

$$P^{\chi}(t_i = 1) = \frac{|\{I : I \in \mathcal{I}_{\chi(u_s)}, t_i = 1\}|}{|\mathcal{I}_{\chi(u_s)}|} \quad . \tag{4}$$

The cue combination for tagging an image I_{u_s} of a seed user u_s can be summarized as follows:

1. Generate the binary ALIPR annotation vector $(a_1, \ldots, a_N)$ for the image I_{u_s}.

2. For each network tag $t_i (i = 1, \ldots, M)$, compute the posterior probabilities of t_i, given ALIPR annotations, i.e.

$$p_i = P^{\chi}(t_i = 1 \mid a_1, \ldots, a_N)$$

Size of ATLST2 Vocabulary	Strategy X	Strategy Y	Alternate Hypothesis H_a
≤3000 (648 networks)	ALL	POP	SNR(X)<SNR(Y)
	POP	ATLST2	SNR(X)<SNR(Y)
	POP	COMPOSITE	SNR(X)<SNR(Y)
	COMPOSITE	**ATLST2**	**SNR(X)< SNR(Y)**
>3000 (242 networks)	ALL	POP	SNR(X)<SNR(Y)
	POP	ATLST2	SNR(X)< SNR(Y)
	POP	COMPOSITE	SNR(X)<SNR(Y)
	COMPOSITE	**ATLST2**	**SNR(X) > SNR(Y)**

Table 1: Paired t-tests for comparing the network vocabulary selection strategies. In all cases, the null hypothesis H_0 : SNR(X)=SNR(Y) was rejected with p-value=0.

and

$$\bar{p}_i = P^{\chi}(t_i = 0 \mid a_1, \ldots, a_N) \ .$$

3. The most likely tag given by this model is the one that has the best odds,

$$t^* = \arg \max_{i=1,\ldots,M} \frac{p_i}{\bar{p}_i} \ . \quad (5)$$

Note that Eq. 5 is an increasing function of p_i and one may approximate it using the value of p_i itself.

4. Since each image typically has multiple correct tags, we rank the M user tags in the decreasing order of the odds, and pick the top K tags as predictions for I_{u_s}, for some predetermined value of K.

5.3 Naive Bayes Formulation

The direct estimation of p_i and $\bar{p}_i$ for each tag t_i is very challenging because of the sparsity issues. If there were a very large number of data instances, then these terms could be estimated. However, for most practical purposes, the volume of the training data on a per-user basis will not be available, and we must do approximate estimation as follows:

$$\begin{aligned} p_i &= P^{\chi}\big(t_i = 1 \mid a_1, \ldots, a_N\big) \quad (6) \\ &= \frac{P^{\chi}(t_i = 1)}{P^{\chi}\big(a_1, \ldots, a_N\big)} P^{\chi}\big(a_1, \ldots, a_N \mid t_i = 1\big) \\ &= \kappa P^{\chi}(t_i = 1) P^{\chi}\big(a_1, \ldots, a_N \mid t_i = 1\big) \ . \end{aligned}$$

where κ is a constant factor independent of t_i. At this point, we make a conditional independence assumption to obtain a Naïve Bayes formulation. The assumption made here is that given the state of a particular tag t_i, the predictions made by ALIPR over its vocabulary are conditionally independent of each other. The equation changes accordingly to

$$p_i = \kappa P^{\chi}(t_i = 1) \prod_{j=1}^{N} P^{\chi}\big(a_j \mid t_i = 1\big) \ . \quad (7)$$

Similarly, the probability for $P^{\chi}\big(t_i = 0 \mid a_1, \ldots, a_N\big)$ is computed as

$$\bar{p}_i = \kappa P^{\chi}(t_i = 0) \prod_{j=1}^{N} P^{\chi}\big(a_j \mid t_i = 0\big) \ . \quad (8)$$

The log odds for t_i are computed by taking a ratio of Eq. 7 to Eq. 8 and converting to the log scale,

$$\begin{aligned} logodds(t_i) = \log P^{\chi}(t_i = 1) + \sum_{j=1}^{N} \log P^{\chi}\big(a_j \mid t_i = 1\big) \\ - \log P^{\chi}(t_i = 0) - \sum_{j=1}^{N} \log P^{\chi}\big(a_j \mid t_i = 0\big) \end{aligned}$$

The log odds can be obtained without knowing κ.

When a tag a_j is predicted by ALIPR, we need an estimate of $P^{\chi}\big(a_j = 1 \mid t_i = 1\big)$ for Eq. 7, for the different values of i. These estimates are sufficient even for cases where a_j is not predicted by ALIPR, since $P^{\chi}\big(a_j = 0 \mid t_i = 1\big) = 1 - P^{\chi}\big(a_j = 1 \mid t_i = 1\big)$. The estimation of the terms $P^{\chi}\big(a_j = 1 \mid t_i = 1\big)$ is given by,

$$P^{\chi}\big(a_j = 1 \mid t_i = 1\big) = \frac{|\{I : I \in \mathcal{I}_{\chi(u_s)}, t_i = 1, a_j = 1\}|}{|\{I : I \in \mathcal{I}_{\chi(u_s)}, t_i = 1\}|} \quad (9)$$

for each $i = 1, \ldots, M$ and $j = 1, \ldots, N$.

Probabilities are smoothed by adding a small count $\epsilon = 1$ to the numerator and the denominator of Eq. 4 and Eq. 9. Also, for practical considerations, only up to 1000 images per member of $\chi(u_s)$ are pooled to build $\mathcal{I}_{\chi(u_s)}$.

6. EVALUATION

Let the proposed tag recommendation framework be denoted as **C+LIN** as it utilizes a combination of visual content and local interaction networks.

For evaluation, we randomly selected 2400 test images from the 890 seed users so that at least two test images per user are covered. The tags given by the seed user to these test images are considered as the ground truth against which the tag predictions are compared to measure performance, both quantitatively and qualitatively.

6.1 Quantitative Evaluation

Quantitative results are presented as the average precision and recall over all test images. For each test image, top $K (K \in \{5, 10\})$ tag recommendations are considered. If P_K is the set of K predictions and G is the ground truth, then,

1. $Precision@K$ - the fraction of correctly retrieved tags among the top K predictions.

$$Precision@K = \frac{|P_K \cap G|}{K} \ .$$

2. $Recall@K$ - the fraction of all actual tags that are

correctly retrieved.

$$Recall@K = \frac{|P_K \cap G|}{|G|} \quad .$$

6.1.1 Baseline Comparison

We compare the results of C+LIN against two baselines: content only (C) and local interaction network only (LIN).

1. **Content (ALIPR) Annotation Baseline (C)**: The top K predictions of ALIPR are used as the first baseline. As mentioned earlier, like any other content based annotation engine, ALIPR's vocabulary is restricted to the concepts used in training which represent only a fraction of the potential user tags. However, this comparison is necessary to set a context for extending content based predictions through folksonomy.

2. **Local Interaction Network Baseline (LIN)**: The second baseline constitutes of the top ranked K tags in the $\tilde{T}_{\chi(u_s)} = \{t_i | i = 1...M\}$.

Fig. 4 presents the results of the quantitative performance comparison with the two baselines showing a significant improvement using the C+LIN combination.

6.1.2 Comparison with Personalization

We also present a scenario where a seed user's information is available, so that personalized tags can be predicted using only the seed user's own tagging behavior in two ways. For both methods, the model training does not include information from the test images.

- **Seed's Popular Tags (S)**: The seed user u_s's vocabulary T_{u_s} has a frequency distribution over her personal tag annotations. We use the top K frequent tags of this vocabulary to annotate the test images. This serves to simulate the user's personal tagging biases and build on idiosyncrasies.

- **Seed's Vocabulary and Content (C+S)**: Instead of the network's vocabulary $\tilde{T}_{\chi(u_s)}$ in the proposed C+LIN framework, we substitute the seed's own vocabulary T_{u_s} and recompute a content based personalized tag recommendation model over the seed user's images $\mathcal{I}_{u_s}$. This serves to simulate the case where the user's personal information is known and is combined with the visual content to predict tags.

The better of the two strategies should be considered as a reasonable upper bound on the performance of the tag recommendation results that can be achieved with the current dataset and the methods studied. In this light, the performance of the proposed system should be evaluated not as an absolute, but in a relative comparison. Fig. 5 shows the comparison of C+LIN with S and C+S.

Table 2 shows the numeric results for precision and recall of all the methods together. It can be seen that the strategy S, where the user's most popular tags are repeatedly applied to the test images has the best performance among all strategies considered so far, with the proposed C+LIN strategy ranking next.

6.2 Qualitative Comparison

The ground truth tags given by the seed user can suffer from folksonomic idiosyncrasies. It may be incomplete (user may choose to not include a relevant tag that would otherwise be found in other similar images) or visually irrelevant (e.g., uncle, zzzzzzz1232zz). Therefore, even some meaningful predictions get penalized. The absolute quantitative performance needs to be coupled with the qualitative assessment of results. In this case, we have provided a few examples for the readers' consideration in Table 3, where the predictions of ALIPR (C), frequency based local network (LIN) and the proposed framework (C+LIN) are presented against the ground truth. For ALIPR, only the predictions with probability greater than 0.1 are shown. For LIN and C+LIN, the top 5 predictions are shown.

The first two images in the first row of Table 3 belong to the same user. Whereas the social network (LIN) provides the same set of tags to both these images and the ALIPR tags may not always be relevant to a user's vocabulary, their combination produces meaningful variations in the user's context. The relevance of these tags is not captured by the the standard quantitative metrics.

	% Precision		% Recall	
	@5	@10	@5	@10
C	2.59	2.14	2.05	3.19
LIN	10.03	7.13	8.12	11.46
C+LIN	**12.02**	**9.57**	**10.85**	**17.40**
S	32.42	23.59	26.25	36.07
C+S	5.33	4.56	5.43	8.66

Table 2: Quantitative Results Summary.

6.3 Effect of Network Size

We found the correlation coefficient for the Recall@10 and the network size to be -0.0995, a statistically significant number with p-value=0.003. This shows that the proposed approach is more effective for small sized local networks and that the performance gradually decreases as the number of connections grow. This is expected owing to the cumulative built up of noisy folksonomic variations as more new connections are incorporated in a network. We believe that more sophisticated statistical approaches need to be incorporated to handle large networks in future.

7. DISCUSSIONS

In this section, we present a few salient observations and arguments from our experiments.

1. The strategy S is the best performing method. This points to the idiosyncratic habit of the users to tag their images similarly, habits which are encouraged by the convenient batch tagging facilities, where users can annotate entire image collections with the same set of tags. Therefore, often in folksonomic settings, a user's previously used tags can be the best determinants of her future uploads.

2. The strategy C+S performs considerably poorly as compared to S as well as C+LIN. This may happen for two reasons. If the user tags are indeed assigned to images in a batch-tag process, then the annotations appear with much diverse (and possibly irrelevant) visual features, which hampers the model learning. Secondly, the user's own tag history may be insufficient to train a useful personalization model. Therefore, we

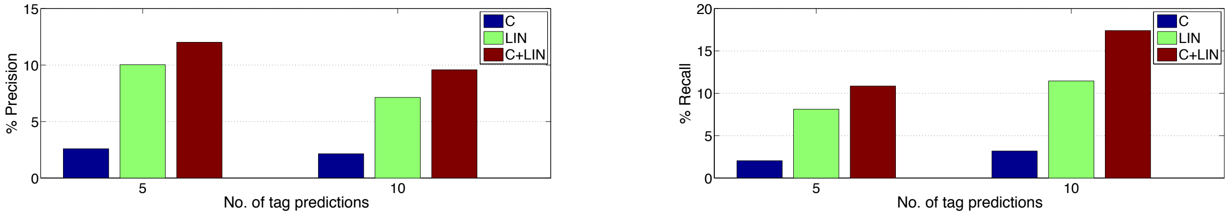


Figure 4: Comparison with baselines.

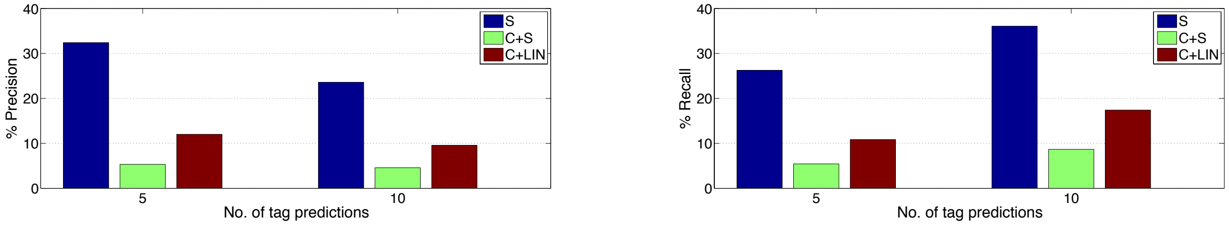


Figure 5: Comparison with personalization strategies.

conclude that even though a user's previous tags may be the best determinants of her future uploads, they may not be useful in learning visual models.

3. On the other hand, we have successfully demonstrated that when the seed user's personal tagging preferences are not known (or when insufficient data is available), a suitable approximation can be made using the tagging behavior of the user's local interaction network. Useful tags can be recommended by combining a substantially smaller tag subset from the local interaction network with the visual concepts detected using a content based annotation method. The combination C+LIN outperforms the baselines of pure content based annotation (C) and the local network (LIN) as well as the model trained using the combination of content and the user's own tags (C+S). This is a key result of our work.

4. As can be seen from the qualitative results, the C+LIN combination can produce interesting and relevant annotations, beyond the tags contained in the ground truth. This suggests that the real performance may be better than what the quantitative tests capture.

8. CONCLUSIONS AND FUTURE WORK

Folksonomic challenges continue to daunt real-world tag recommendation systems. In this paper, we have experimented with the idea of exploiting a user's local interaction network for automatically tagging her photos. We have looked at its performance in isolation as well as in conjunction with a content-based image annotation system. By experimenting on a large, real-world dataset from Flickr, we have been able to draw inferences about the effectiveness of the proposed approach. Our observation has been that the use of the local interaction network is quite effective at boosting annotation performance, though performance may be improved in principal, if a user's own idiosyncrasies are known. The use of networks can capture the correlation between the visual concepts and the textual tags better and set a context for the seed user. By combining the social network cues with the visual content, we have also been able to extend the capabilities of the traditional content based tag recommendation systems to handle a potentially unlimited vocabulary and provide annotations that are meaningful in the context of a user and her social network, even when the user's own tagging preferences are not known.

This work presents our first step in using the local networks for user characterization and tag recommendation. We realize that the social networks are often complex and can be characterized using many different properties. We believe that understanding and incorporating these properties intelligently, can further boost user characterization in a way that can be used to support interesting applications, not only image-related but also relevant in other social networking domains.

9. ACKNOWLEDGMENTS

This work is supported in part by the US National Science Foundation under Grant Nos. IIS-0347148, CNS-0202007, DMS-0705210 and by The Pennsylvania State University. We acknowledge Flickr API for the test data.

10. REFERENCES

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

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Tagging over time: Real-world image annotation by lightweight meta-learning. In *Proc. ACM Multimedia*, pages 393–402, 2007.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[4] Facebook. http://www.facebook.com.

[5] Facebook News. http://www.facebook.com/note.php?note_id=30695603919.

[6] Flictr Blog. http://blog.flickr.net/en/2008/11/03/3-billion.

[7] N. Garg and I. Weber. Personalized, interactive tag recommendation for flickr. In *Proc. ACM Recommender Systems*, pages 67–74, 2008.

[8] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Information Science*, 32(2):198–208, 2006.

[9] Google Picasa. http://picasaweb.google.com.

[10] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. ACM Special Interest Group on Information Retrieval*, pages 119–126, 2003.

[11] O. Kucuktunc, S.G. Sevil, A. Tosun, H. Zitouni, P. Duygulu, and F. Can. Tag suggestr: Automatic photo tag expansion using visual information for photo sharing websites. In *Proc. Semantic and Digital Media Technologies: Semantic Multimedia*, pages 61–73, 2008.

[12] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[13] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[14] X. Li, L. Chen, L. Zhang, F. Lin, and W.-Y. Ma. Image annotation by large-scale content-based image retrieval. In *Proc. ACM Multimedia*, pages 607–610, 2006.

[15] S. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, and G. Thallinger. Automatic image annotation using visual content and folksonomies. *Multimedia Tools and Applications*, 42(1):97–113, 2009.

[16] S. Lindstaedt, V. Pammer, R. Mörzinger, R. Kern, H. Mülner, and C. Wagner. Recommending tags for pictures based on text, visual content and user context. In *Proc. Internet and Web Applications and Services*, pages 506–511, 2008.

[17] E. Moxley, J. Kleban, and B.S. Manjunath. Spirittagger: a geo-aware tag suggestion tool mined from flickr. In *Proc. ACM Multimedia Information Retrieval*, pages 24–30, 2008.

[18] M. Naaman and R. Nair. Zonetag's collaborative tag suggestions: What is this person doing in my phone? *IEEE MultiMedia*, 15(3):34–40, 2008.

[19] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *Proc. Joint Conference on Digital Libraries*, pages 178–187, 2005.

[20] R. Negoescu and D. Gatica-Perez. Analyzing flickr groups. In *Proc. Content-based Image and Video Retrieval*, pages 417–426, 2008.

[21] Orkut. http://orkut.com.

[22] Photobucket. http://photobucket.com.

[23] S. Sen, S. Lam, A. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. Harper, and J. Riedl. Tagging, communities, vocabulary, evolution. In *Proc. Computer Supported Cooperative Work*, pages 181–190, 2006.

[24] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proc. Recommender Systems*, pages 259–266, 2008.

[25] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. World Wide Web*, pages 327–336, 2008.

[26] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[27] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *Proc. Computer Vision and Pattern Recognition*, pages 1483–1490, 2006.

[28] Yahoo! Flickr. http://www.flickr.com.

[29] Yahoo! Image Search. http://images.search.yahoo.com.

[30] R. Yan, A. Natsev, and M. Campbell. An efficient manual image annotation approach based on tagging and browsing. In *Proc. Multimedia Information Retrieval Workshop, ACM Multimedia*, pages 13–20, 2007.

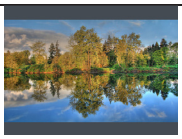
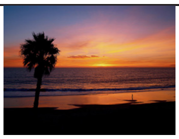
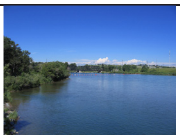
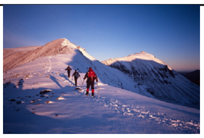
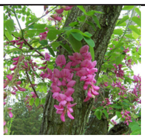
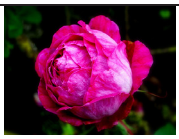
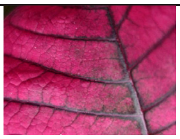
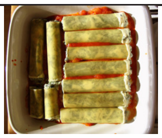
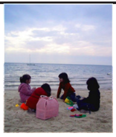
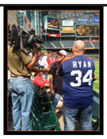
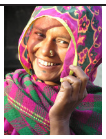
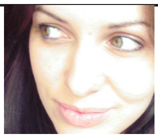
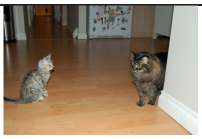
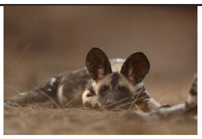
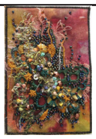
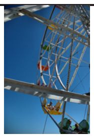
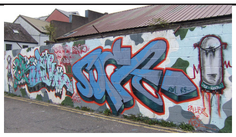
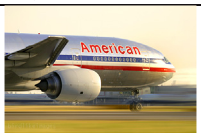
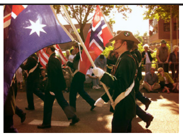
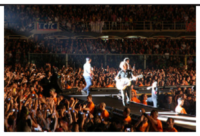
				
Ground truth	bravo, pond, trees, oregon, canon	beach, naturescenes, sunset, wow, orange	canada, bowriver, calgary, tree, alberta	scotland, walking, mountain, winter
C (ALIPR)	landscape	indoor, modern, sky, sunset, sun	sky, ocean, wild_life, bird, people	ocean, boat, water, sky, ski
LIN	water, red, sunset, abigfave, flower	water, red, sunset, abigfave, flower	water, sky, blue, reflection, sunset	scotland, sea, mountain, cycling, coast
C+LIN	nature, bravo, nikon, flower, sky	silhouette, sunset, dawn, topv111, sunrise	sky, canada, alberta, blue, drumheller	ice, snow, scotland, highlands, switzerland
				
Ground truth	flower, flowers, trees, fragrant, tree	flower, bud, rose, pink, macro	bloom, flower, buds, winter	italian, food
C (ALIPR)	flower, plant, house	flower, plant, orchid	flower, pattern, natural, cloth, indoor	food, indoor, cuisine, man-made
LIN	flower, sunset, blue, red, sky	water, flower, sunset, red, green	flower, green, yellow, macro, flowers	groningen, netherlands, nederland, green, water
C+LIN	garden, flowers, nature, flower, squirrel	flower, flora, rosebud, timepiece, yellow	rose, macros, closeup, interestingness, flower	salad, sushi, food, chips, japan
				
Ground truth	girls, kids	baseball, houstonastros, minutemaidpark, houston	india, portrait, smile, closeup, faces	smile, lips, eyes. face
C (ALIPR)	indoor, manmade	manmade	people	indoor, drawing
LIN	green, sky, blue, canon, sunset	houston, texas, vacation, tx, landscape	india, sky, light, blue, red	girl, clouds, lake, portrait, art
C+LIN	boy, baby, kids, children, child	houston, texas, slightclutterphotography, sanantonio, tx	india, red, d70, life, people	photoshop, art, women, white, black
				
Ground truth	cats, cat, kitten, kittens	safari, africanwilddog. nature, africa, wildlife	postcard, embroidery, fabricpostcard, sewing, beads	california
C (ALIPR)	animal, wild_life	animal, wild_life	man-made	ocean, people, water
LIN	cat, bestofcats, kissablekat, impressedbeauty, cats	nature, anawesomeshot, bird, sunset, impressedbeauty	embroidery, art, fabric, blue, flower	hotel, reflection, sunset, river, fountain
C+LIN	cat, kissablekat, pet, kitty, tabby	specanimal, nikon, birds, wildlife, africa	handmade, embellisher, fabric, art, embroidery	bridge, california, buildings, flag, sanfrancisco
				
Ground truth	art, ireland, cork, graffiti, streetart	airplane, heathrow, 777, jet, boeing	flags, flag, parade, army	italy, luce, florence, firenze, musica
C (ALIPR)	indoor, man-made	indoor	people, cloth	man-made
LIN	graffiti, art, street, ireland, cork	sunset, sky, blue, water, night	path, mosaic, anawesomeshot, hill, bokeh	italy, firenze, italia, florence, blueribbonwinner
C+LIN	ireland, limerick, art, fun, street	aviation, airplane, aircraft, plane, airport	parade, nikon, presidentialprimary, elections, candidates	italy, italia, abigfave, firenze, florence

Table 3: Qualitative Results