Contextual Wisdom: Social Relations and Correlations for Multimedia Event Annotation

Amit Zunjarwad *Hari Sundaram* *Lexing Xie*

Arts, Media and Engineering, Arizona State University IBM TJ Watson Research Center

Email:{amit.zunjarwad, hari.sundaram}@asu.edu xlx@us.ibm.com

ABSTRACT

This work deals with the problem of event annotation in social networks. The problem is made difficult due to variability of semantics and due to scarcity of labeled data. Events refer to real-world phenomena that occur at a specific time and place, and media and descriptions are treated as facets of the event metadata. We are proposing a novel mechanism for event annotation by leveraging related sources (other annotators) in a social network. Our approach exploits event concept similarity, concept co-occurrence and annotator trust. We compute concept similarity measures across all facets. These measures are then used to compute event-event and user-user activity correlation. We compute inter-facet concept co-occurrence statistics from the annotations by each user. The annotator trust is determined by first requesting the trusted annotators (seeds) from each user and then propagating the trust amongst the social network using the biased PageRank algorithm. For a specific media instance to be annotated, we start the process from an initial query vector and the optimal recommendations are determined by using a coupling strategy between the global similarity matrix, and the trust weighted global co-occurrence matrix. The coupling links the common shared knowledge (similarity between concepts) that exists within the social network with personalized observations (i.e. concept co-occurrences) that the user trusts. Our initial experiments on annotated everyday events are promising and show substantial gains against traditional SVM based techniques.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**] Information filtering, search process

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Social networks, context, event annotation, images, content management, multimedia

1 INTRODUCTION

In this paper we address the problem of event centric image annotation by exploiting activity correlation amongst members of a social network within a trusted context. The paper makes possible for members of a social network to effectively annotate images. The problem is important as online media sharing sites such as Flickr [1], are enormously popular (and more recently event centric media sites such as SEraja [2]) and yet since tags are still rare or sparse. In such systems, text is predominantly used to search for media, and the absence of robust annotations will preclude effective search. Developing robust concept classifiers for annotation of everyday images is very challenging problem, due to visual diversity of such images and tag scarcity.

Why should a social network be helpful for image annotation? The key observation is that members of the social network have highly correlated real-world activities – i.e. they will participate in common activities together, and often repeatedly. We conjecture that the shared activities provides valuable context for both the images and the descriptions – they will be of shared events, images of mutual friends and the semantics of the tags used to describe them will be consistently understood within the social network.

The identification of correlated (in the sense of activity semantics) members within the social network may lead to higher quality recommendations due to the pooling of observations (images and text tags) of the correlated members. From a pattern recognition point of view, the ability to pool images essentially increases the ground truth available per tag. It has the important benefit that the tags are likely to be used in the same context (i.e. share the semantics) as the user activity is correlated.

1.1 Related work

There has been recent interest in '*folksonomy*' [19,23]. It has been noted that a large number of ordinary untrained folk are tagging media as part of their everyday encounters with the web (http://del.icio.us), or with media collections [1]. The attraction of folksonomy lies in the idea that collective tagging can significantly reduce the time to determine media that are semantically relevant to the users, for example as part of a search.

There has been prior work in using groups for the purposes of image annotation / labeling [3,22]. In the ESP game [3], the authors develop an ingenious online game, in which people play against each other to label the image. In [22] the authors take into account browsing history with respect to an image search for determining the sense associated with the image. Both work aims at recovering one *correct* sense either shared by common knowledge or the user's own history. The context in which the annotation is used / labeled is not taken into account. In [25] the authors explore a collaborative annotation system for mobile devices. There they used appearance based recommendations as well as location context to suggest annotations to mobile users. In [16], the authors provide label suggestions for identities based on patterns of re-occurrence and co-occurrence of different people in different locations and events. However, they do not make use of user-context, or commonsensical and linguistic relationships and group semantics.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

In [6,13], the authors use sophisticated classification techniques for image annotation. However, they do not investigate collaborative annotation within a social network. The image based classifier schemes run into two broad problems: (a) scalability – each tag, requires its own classifier, and (b) the fact that people may use a tag in very different senses makes the classifiers difficult to build.

While there has been very little work on the role of trust for image annotation, there has been work on trust in the context of web-spam detection [9]. In this paper the authors develop an algorithm that that propagates trust from a small set of seed pages evaluated by an expert. Their main intuition is that good pages rarely point to bad pages. The web spam detection problem has similarities to image annotation in terms of cost. Today, spam is detected manually – it is expensive, but extremely important for web search engines to be able to filter out such sites to ensure high quality search results. Our work on annotator trust has been motivated by their problem formulation.

A key limitation of prior work on image annotation is that there is an implicit assumption that there is one correct semantic associated with the image that needs to be uncovered by classification. In social networks the assumption of consistent labeling of images (thus implying semantic agreement) over the dataset may not hold over a diverse set of concepts. In prior work [17], we have observed that there is non-negligible disagreement among users, particularly on concepts that are more abstract rather than concrete. For example, people are more likely to disagree on abstract concepts such as "love", "anger", "anxiety" etc. as compared to everyday concepts such as "pen", "light bulb", "ball" etc. The implication is that building a concept classifier to annotate an image, will not work well across *all* users, particularly for abstract concepts.

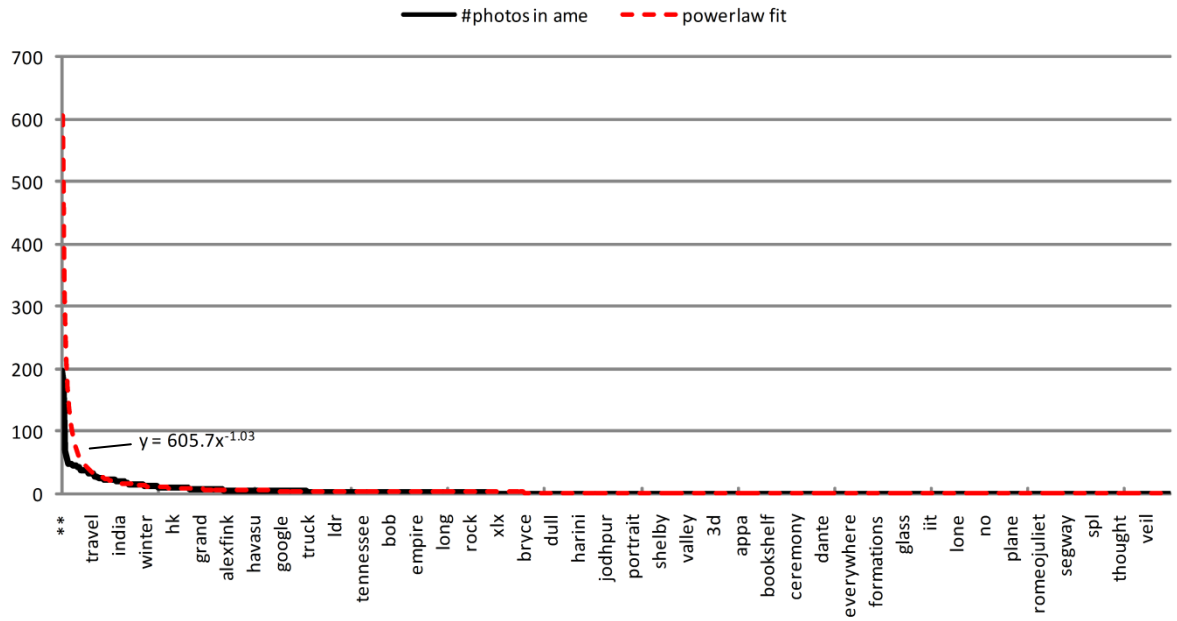

Figure 1: The AME pool on Flickr, showing the familiar power-law distribution of the tags. The power law equation is $y = 605.7x^{-1.03}$.

1.2 Our Approach

We are proposing an event-centric approach to media annotation, which incorporates social network trust. Specifically, we do not develop per-concept classifiers, to address the classifier scalability issue. Instead our approach to annotation is motivated by web search algorithms such as HITS [12] and PageRank [5]. In these algorithms, query-relevant documents are found through iterative mechanisms on the hyperlinked structure, instead of *pre-classifying documents using concept classifiers.*

Our approach is grounded in observations of the tag distribution in the Flickr dataset (ref. Figure 1). We observe that the tag distribution follows the familiar power law distribution found in online social networks [4,20]. These observations have consequences for concept based annotation systems, in terms of learnability (not enough data for most tags), scalability (too many classifiers – this will become computationally expensive) and semantic variability (due to different user contexts).

We define events to be a real-world occurrence, which may be described using attributes such as images, and facets such as who, where, when, what. A key idea is that media (including images and text) are event meta-data – i.e. they are description of the event, not the event itself [24]. We refer to event descriptions via the attributes of images and text as the event context – *these set of attributes / facets that support the understanding of everyday events*.

Given a social network and events, we compute event concept similarity, concept co-occurrence and annotator trust. We compute concept similarity measures across all facets (who, where, when what and image) using ConceptNet [14] as well as low-level features. These measures are then used to compute event-event and user-user activity correlation. We compute inter-facet concept co-occurrence statistics from the annotations by each user. The annotator trust is determined by first requesting the trusted annotators (seeds) from each user and then propagating the trust amongst the social network using the biased PageRank algorithm.

The recommendation algorithm is a variant of the well known HITS algorithm [12]. The optimal recommendations are determined by using a coupling strategy between the global

similarity matrix, and the trust weighted global co-occurrence matrix. The trust is computed for each user, over the entire social network. The coupling links the common shared knowledge (similarity between concepts) that exists within the social network with personalized observations (i.e. concept co-occurrences) that the user trusts. Our preliminary experimental results when compared to traditional concept classifiers are promising.

The rest of this paper is organized as follows. In the next section we present some observations relating to image annotation. In Section 3, we introduce the idea of events. In Section 4, we discuss inter-facet distance and co-occurrence statistics. In Section 5, we develop the idea of annotator trust and follow that section with a section on generating concept recommendations for images. In Section 7, we present our experimental results, and then conclude the paper with our summary and conclusions.

2 AN ANNOTATION PUZZLE

In this section, we present some observations and challenges that occur in annotating images from everyday events.

2.1 The long Tail

The "long tail" [4] refers to a power law distribution of entities observed in online problem domains where large groups of people interact. We now present observations from a community pool in Flickr as well as statistics from Flickr.

We begin with an analysis of the AME (Arts, Media and Engineering Program, the home institution of the first two authors) flickr pool. The pool distribution at the time of writing the paper (~1200 photos, ~575 unique tags, 41 members), shows the familiar power law distribution ($y = 605.7x^{-1.03}$, ref Figure 1). An interesting observation is that only about 11% of the tags (67 / 575) contain more than 10 photos. Furthermore the top two tags are both names of a pool member, who by habit tags *all of her photos* by variants of her name.

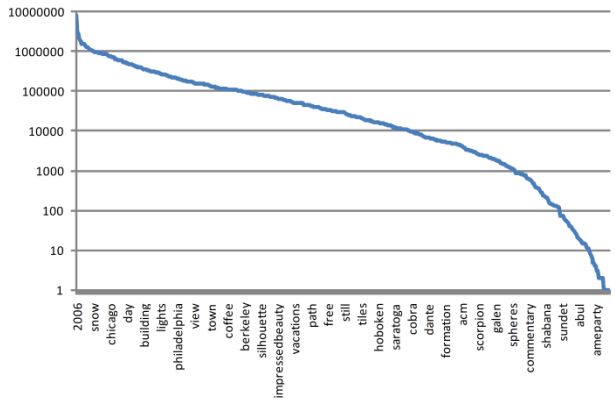

Figure 2: Flickr global pool distribution for the same tags as the AME pool. The figure shows that most tags occur frequently – note that the vertical axis is a logarithmic scale.

We now examine the global Flickr pool of the same tags as in the AME pool (ref. Figure 2). The figure shows the frequency of occurrence (the vertical axis is logarithmic, for the sake of clarity) of the same tags in the global flickr pool. The data shows that most tags (~90%, 522/575) have at least 100 photos associated with them. We note that the global pool is data is not a power-law distribution because it is not a plot of *all the global tags*. Furthermore most of the photographs for the frequently occurring tags (e.g. 2006, trip, china) are highly visually diverse.

2.2 Three Problems

The long tailed distribution image tags on the community AME Flickr pool, and in Flickr in general, raises important questions training classifiers for generating recommendations. Specifically, there are concerns relating to concept learnability, classifier scalability and the role of context in concept learning.

2.2.1 Concept Learnability

The AME and global flickr pools show very different numbers of photos tagged with the same word. Given that most of the tags in the community have very few photos instances (only 10% have more than 10 photos), it is very difficult to learn concept classifiers for most of the tags. While many of the tags in the AME do have many more positive instances in the global pool, they can be highly visually diverse, thus making the classifier weak.

The main issue here is that the power-law distribution is fundamental characteristic of large datasets from online communities such as Flickr [4]. Then, the consequence of this observation is that most of the tags even in the global pool will have very few positive instances due to the power-law characteristic. This makes the construction of concept classifiers difficult, for most tags.

2.2.2 Classifier Scalability

Classifier scalability deals the issue of *number* of useful classifiers. While there has been attempts to develop a multimedia ontology (LSCOM [11], 449 concepts) for domains such as news video, this is a challenging problem in unstructured domains such as photographs from everyday events. Flickr has an extremely large number of tags, and learning a global concept classifier for each unique tag makes the automated image annotation problem computationally expensive. This is because we would need to test each trained classifier on the untagged image. We note that even in the AME group pool there are a total of 575 tags for only ~1200 photographs.

2.2.3 The role of Context

Learning a classifier on a set of images tagged with the same keyword implicitly assumes that the photos share the same context in which the keyword is appropriate. An examination of both the AME flickr pool and the Global pool reveals that this is not accurate. For example in the AME flickr pool, there are photos tagged as "saguaro" – the photos exist in two contexts – the cactus, and the name of a lake. In the flickr global pool, there are thousands of photographs labeled as "yamagata" – some are of the town, some refer to the visual artist (Hiro Yamagata), while still others refer to the singer (Rachel Yamagata).

What is missing from both the AME and the global flickr pools is the *context* in which tag makes sense for the author / annotator of the photo. This lack of context makes it difficult to use one classifier per concept trained on all photographs in the Flickr pool, on a photograph whose context is not known.

We are proposing an event-centric approach to media annotation, which incorporates social network trust. Specifically, we do not develop per-concept classifiers, to address the classifier scalability issue. Instead our approach to annotation is motivated by web search algorithms such as HITS [12] and PageRank [5]. We next discuss the notion of an event.

3 WHAT ARE EVENTS?

In this section we provide a formal definition of events and introduce the idea of the event context.

3.1 Definition

An event refers to a real-world occurrence, which may be described using attributes such as images, and facets such as who, where, when, what. Events may be spread over temporal and spatial attributes. For example an event "new year's eve celebration" can occur at multiple locations and at different times (due to time-zone differences). Events such as "John's party" may take place at a single location, but may be spread over a few hours. Events may also have temporal structure – "Weekly Lunch discussion with Mary" etc. In this paper we have restricted our focus to events that occur over a single location and contiguous time – we do not consider event hierarchies or event temporal structures. This was done for computational simplicity.

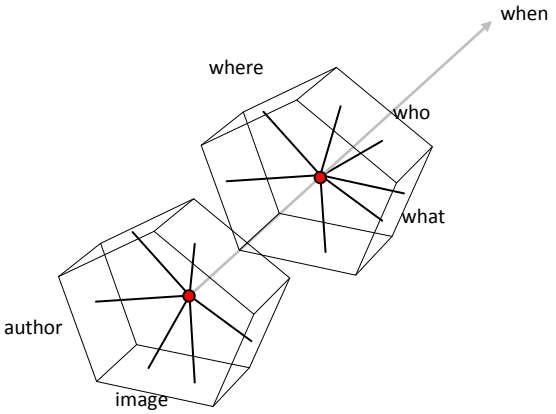

Figure 3: The figure shows two events (the two red dots) along a timeline, where each event is shown to last a contiguous period in time. The first event has two images associated with it, while the second event has two words in the "what" facet.

Our understanding of events draws upon recent work by Jain and Westermann [24]. A key idea in that paper was the notion that media (including images and text) are event meta-data – i.e. they are description of the event, not the event itself. This is a reversal of the traditional relationship between media and events, where media (e.g. video / images) contain the event to be found. The Jain-Westermann approach suggests that media contain partial descriptions of the real-world event, and these descriptions need to be gathered to develop a full understanding of the event. A consequence of adopting this idea is that in our framework, events can contain multiple text tags, as well as multiple images, all of whom describe the event.

3.2 Event Context

We refer to event descriptions via the attributes of images and text as the event context – *these set of attributes / facets that support the understanding of everyday events*.

The notion of "context" has been used in many different ways across applications [7]. Note that set of contextual attributes is always application dependent [8]. For example, in ubiquitous computing applications, location, identity and time are critical aspects of context [7]. In describing everyday events the *who*, *where*, *when*, *what* are among the most useful attributes, just as basic journalism would teach "3w -- who when where" as the basic background context elements for reporting any real-world event.

4 SIMILARITY AND CO-OCCURRENCE

In this section we present our approach to computing the similarity between any two concepts along a specific facet, as well as the role of the inter-fact co-occurrence matrix. Both similarity and co-occurrence are then used to compute the recommendations.

Similarity and co-occurrence represent different forms of knowledge, which are used in our system. Similarity measures typically represent "global knowledge" used by the algorithm designer to address the content analysis problem – they are user independent. The co-occurrence matrix represents personal knowledge – i.e. assertions about two facts (e.g. where = "new york", what = "fun") that are useful perhaps only to one individual. Of course these assertions may also encode assertions that are widely shared.

4.1 Intra-Facet Concept Similarity

We now discuss the similarity measures for the different event facets. We first derive a new ConceptNet based event similarity measure for a pair of concepts. We then extend this similarity measure to two sets of concepts. Similarity measures over the context facets are then defined using the two above measures.

4.1.1 The ConceptNet based semantic distance

In this section, we shall determine a procedure to compute semantic distance between any two concepts using ConceptNet – a popular commonsense reasoning toolkit [14].

ConceptNet has several desirable characteristics that distinguish it from the other popular knowledge network – WordNet [15]. First, it expands on pure lexical terms to include higher order compound concepts ("buy food"). Secondly, it greatly expands on three relations found in WordNet, to twenty. The repository represents semantic relations between concepts like *"effect-of"*, *"capable-of"*, *"made-of"*, etc. Finally, ConceptNet is powerful because it contains practical knowledge – it will make the association that "students are found in a library" whereas WordNet cannot make such associations. Since our research is focused on recommending annotations to images from everyday events, ConceptNet is very useful.

The ConceptNet toolkit [14] allows three basic functions on a concept node [14]:

- `GetContext(node)` – this finds the neighboring relevant concepts using spreading activation around the node. For example – the neighborhood of the concept *"book"* includes *"knowledge"*, *"library"*, *"story"*, *"page"* etc. ConceptNet terms this operation as "contextual neighborhood" of a node.
- `GetAnalogousConcepts(node)` – Two nodes are analogous if they derive incoming edges (note that each edge is a specific relation) from the same set of concepts. For example – analogous concepts for the concept *"people"* are *"human"*, *"person"*, *"man"* etc.
- `FindPathsBetweenNodes(node1,node2)` – Find paths in the semantic network graph between two concepts, for example – path between the concepts

"apple" and *"tree"* is given as *apple [isA] fruit, fruit [oftenNear] tree*.

Neighbors of Concepts: Given two concepts *e* and *f*, the system determines all the concepts in the contextual neighborhood of *e*, as well as all the concepts in the contextual neighborhood of *f*. Let us assume that the toolkit returns the sets C_e and C_f containing the contextual neighborhood concepts of *e* and *f* respectively. The context-based semantic similarity $s_c(e,f)$ between concepts *e* and *f* is now defined as follows:

$$s_c(e,f) = \frac{|C_e \cap C_f|}{|C_e \cup C_f|},$$

<1>

where $|C_e \cap C_f|$ is the cardinality of the set consisting of common concepts in C_e and C_f and $|C_e \cup C_f|$ is the cardinality of the set consisting of union of C_e and C_f.

Analogous Concepts: Given concepts *e* and *f* the system determines all the analogous concepts of concept *e* as well as concept *f*. Let us assume that the returned sets A_e and A_f contain the analogous concepts for *e* and *f* respectively. The semantic similarity $s_a(e,f)$ between concepts *e* and *f* based on analogous concepts is then defined as follows:

$$s_a(e,f) = \frac{|A_e \cap A_f|}{|A_e \cup A_f|},$$

<2>

where $|A_e \cap A_f|$ is the cardinality of the set consisting of common concepts in A_e and A_f and $|A_e \cup A_f|$ is the cardinality of the set consisting of union of A_e and A_f.

Number of paths between two concepts: Given concepts *e* and *f*, the system determines the path between them. The system extracts the total number of paths between the two concepts as well as the number of hops in each path. The path-based semantic similarity $s_p(e,f)$ between concepts *e* and *f* is then given as follows:

$$s_p(e,f) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_i},$$

<3>

where N is the total number of paths between concepts *e* and *f* in the semantic network graph of ConceptNet and h_i is the number of hops in path *i*.

The final semantic similarity between concepts *e* and *f* is then computed as the weighted sum of the above measures. We use equal weight on each of the above measures (in the absence of a strong reason to support otherwise), and write the concept similarity *CS* the as follows:

$$CS(e,f) = w_c s_c(e,f) + w_a s_a(e,f) + w_p s_p(e,f),$$

<4>

where $w_c = w_a = w_p = 1/3$.

In the next subsections, we use ConceptNet distances to compute distances in the *where* and *what* facets of the user and event context, since these two facets are described with a free-form natural vocabulary on which ConceptNet similarities are meaningful, while other facets such as *who* and *when* use quantitatively distances on time, or intersection on proper nouns.

4.1.2 Similarity between two sets of concepts

An event usually contains a number of concepts in a facet; therefore we also need a similarity measure between sets of concepts based on that between two individual concepts. We define the set similarity between two sets of concepts A and B, where A: $\{a_1, a_2, \ldots\}$ and B: $\{b_1, b_2, \ldots\}$, given a similarity measure *m(a,b)* on any two set elements *a* and *b* in the following manner.

$$S_H(A,B \mid m) = \frac{1}{|A|}\sum_{k=1}^{|A|}\max_i \{m(a_k,b_i)\},$$ <5>

This is the average of the maximum similarity of the concepts in set A with respect to the concepts in set B, where |A| is the cardinality of set A. The equation indicates that the similarity of set A with respect to set B is computed by first finding the most similar element in set B, for *each* element in set A, and then averaging the similarity scores with the cardinality of set A. S_H is a variant of the familiar Hausdorff point set distance measure used to compare sets of image features [10] from which we adapt for measuring similarity. We average the similarity instead of using the `min` operator as used in the original Hausdorff distance metric, since averaging is less sensitive to outliers. Like the original Hausdorff distance metric, this similarity measure is asymmetric with respect to the sets: $S_H(A,B|s) \neq S_H(B,A|s)$.

4.1.3 Similarity across event attributes

We now briefly summarize the similarity measures used for each attribute of an event. This is useful in determining if one event is similar to another, as well as user to user similarity. Let us assume that we have two events e_1 and e_2. Note that measures are asymmetric and *conditioned on event* e_2.

- **what**: The similarity in the *what* facet is given as:

$$s(A_1,A_2) = S_H(A_1,A_2 \mid CS),$$ <6>

where A_1 and A_2 refer to the sets of concepts for the *what* facets of events e_1 and e_2 respectively.

- **who**: The similarity $s(P_1,P_2)$ for the *who* facet is defined as:

$$s(P_1,P_2) = \frac{|P_1 \cap P_2|}{|P_2|},$$ <7>

where p_1 and p_2 are the set of annotations in the *who* facet of events e_1 and e_2.

- **where**: The similarity $s(l_1, l_2)$ for the *where* facet is given as:

$$s(L_1,L_2) = \frac{1}{2}\left(\frac{|L_1 \cap L_2|}{|L_2|} + S_H(L_1,L_2 \mid CS)\right),$$ <8>

Where L_1 and L_2 refer to the sets of concepts for the "location" facets of events e_1 and e_2 respectively The equation states that the total similarity between L_1 and L_2 is the average of the exact location intersection with the modified Hausdorff similarity.

- **when**: The similarity $s(t_1,t_2)$ for the *when* facet is given as:

$$s(t_1,t_2) = \frac{1}{2}\left(\frac{|t_1 \cap t_2|}{|t_2|} + S_H(t_1,t_2 \mid CS)\right),$$ <9>

where t_1 and t_2 are the event time *text annotations*, for the time facets of events e_1 and e_2 respectively. Since we are building an event annotation system, we wished to provide textual annotation such as "holidays" to describe the time of the event.

We found in our preliminary experiments with users that they preferred this mode of notating time, rather than the time of the photo. Such annotations allowed them to describe time qualitatively (e.g. "happy"). Note that the time that the photograph was taken can be trivially obtained from the EXIF data of the image, and added as an annotation.

- **Image:** In our work, the feature vector for images comprises of color, texture and edge histograms. The color histogram comprises of 166 bins in the HSV space. The edge histogram consists of 71 bins and the texture histogram consists of 3 bins. We then concatenate these three histograms with an equal weight to get the final composite feature vector. We then use the Euclidean distance between the feature histograms as the low-level distance between two images.

The event similarity measure (ES) between two events can then be defined as a weighted sum of the similarity measures across each event attribute.

$$ES(e_1, e_2) = \sum_{i=1}^{5} \omega_i s(e_1, e_2; i) \quad <10>$$

Where, s_i is the similarity measure of each attribute described in the preceding paragraph and ω_i is the weight of each similarity measure.

4.1.4 The global similarity matrix

We now show how to compute the global similarity matrix $\mathbf{M}_s$. Typically users will have varied annotations for all the event facets. Then:

1. The dimensionality of the global matrix is determined by determining the number of unique attribute values across *all users.* Thus we have all the concepts, per facet that are in use in the social network.
2. We compute similarity values only within each facet. The similarity is computed according to the specific formula for that facet. Each facet's similarity matrix then creates a sub-matrix within $\mathbf{M}_s$.
3. The matrix $\mathbf{M}_s$ is row-normalized, such that each row sums to unity. It is easy to see that the global similarity matrix is block-diagonal, with each block corresponding to the similarity sub-matrix from each facet.

In this paper, the similarity between any two concepts forms a shared universal knowledge amongst all the people annotating events – i.e. we assume that the similarity values are shared. This is a simplifying assumption, and is used here for computational efficiency reasons. For example in [18] we show people can show semantic disagreement over the same image. We acknowledge that a more sophisticated system that allows for a personalized similarity measure would be very useful in this problem. We next discuss the computation of the global co-occurrence matrix.

4.2 Co-occurrence

This section presents our approach to exploiting inter-facet co-occurrence. We cannot compute the similarly between two terms that appear in different facets. For example, if an event is annotated as "home" (where) and "John," (who) then the notion of similarity between these terms is not very meaningful. What we can calculate is the joint probability of any two terms, given all the event annotations of a single person. Note that the joint concept probability distribution can be different across users.

Co-occurrence can reveal personalized associations. Through the analysis of the joint distribution, we can determine highly specific, personalized associations – for example, if a user associated "business trip" (what) with "New York" (where), then for images that are likely to be labeled as New York, we should also recommend "business trip." Clearly, the associations can differ across people.

The concept co-occurrence matrix $\mathbf{M}^k_c$ is computed separately for each user *k*. Let us assume that the user has annotated an event with N concepts (spread over the facets who, where, when and what, and the event label). Then we have N^2 pairs of concepts. The frequency count of each pair in the matrix $\mathbf{M}^k_c$ is then incremented by 1. In practice the co-occurrence matrix is sparse.

The global co-occurrence matrix $\mathbf{M}_c$ is computed by using the co-occurrence matrixes of all the users. Typically users will have varied annotations for all the event facets. This implies that the *dimensions* of the each user's co-occurrence matrix may have unique attribute values. Then:

4. The dimensionality of the global matrix is determined by determining the number of unique attribute values across *all users.*
5. The frequency value for any element (*i* , *j*) of the global matrix is obtained by aggregating the number of observations across all users for the same tuple in their personal co-occurrence matrixes. For example, to compute the global frequency count of the tuple (where = "New York", and what = "business trip"), we need to find all photos for all users what that have tagged their photos with this pair.

$$\mathbf{M}_c \big| (a,b) = \sum_{i=1}^{k} \mathbf{M}^k_c \big| (a,b) \quad <11>$$

Where *a* and *b* represent attribute values (e.g. "New York" and "business trip"). The equation states that the frequency value of the global matrix, subject to the logical predicates (*a*, *b*), is the sum of the frequency counts over all individual user matrixes for the same predicate. Note that if for some user *k*, the predicate does not hold true, then the corresponding user will not contribute to the value of this cell.

In this section we examined two different forms of knowledge – similarity (global) and co-occurrence (personal). We showed how intra-facet similarity can be computed using ConceptNet and low-level feature similarity. Then we determined the co-occurrence matrix per person, and then showed how to create a global co-occurrence representation. We next explore the idea of social network trust.

5 SOCIAL NETWORK TRUST

In this section we show how to determine the trust distribution over a single user's social network. The trust vectors are different for each person in the network.

It must be emphasized that the word "trust" is used in a narrow interpretation here. We clarify this issue, since the word "trust" has very broad semantic connotations. If there are two users, John and Mary and Mary can provide high quality annotations

for John's photographs, then we say that "Mary is a trustworthy annotator" of John's photographs.

5.1 Trusted Social Network Context

Our approach to determining trust exploits both *a priori* knowledge from and data driven activity correlation.

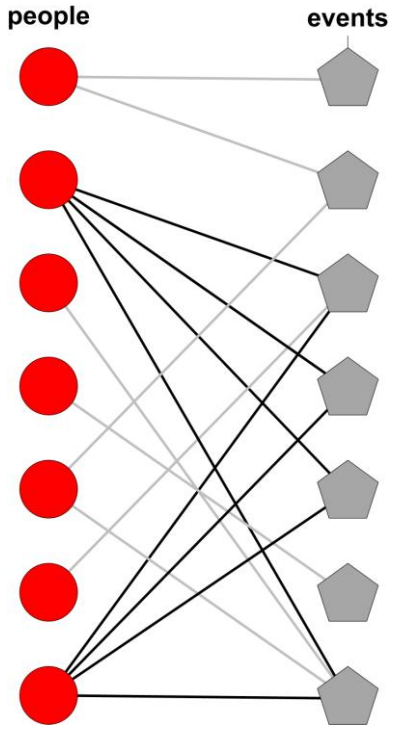

Figure 4: The figure shows how activity correlation can occur between members of the same social network, across events. The second and the last member are highly correlated (black lines used for emphasis.)

A user may have identified certain members of her social network that she trusts as good recommenders – for example, a person may have her spouse as the recommender. This trust cannot be easily inferred from the data, as the person-person relationship may be hidden in the image meta-data. Secondly, it is difficult to place a trust value on a specific relationship type – i.e. even if a photo tag suggests a specific relationship between two people (e.g. father-son), it is difficult to determine the trust value for this relationship.

Trust is real valued number between 0 and 1. However, typically users prefer to specify either 0 or 1 for each member of their social network, suggesting either no trust, or complete trust. Hence the *a priori* trust vector for any user, over the entire network is binary valued.

We compute an *activity based trust* between two users. The main idea is that if two people are highly correlated in terms of their real-world events, then this correlation has an effect on the event descriptors. In Figure 4, we show a sample social network. The red circles represent people, who participate in events (pentagons). The edges represent participation of a person in a specific event. The schematic shows that there exist two users whose show high activity correlation.

In earlier work on image annotation [18], we showed that the members who belong to the same social network tend to agree more with each other than when the members did not belong to a social network. Hence by doing a data driven analysis of the event annotations for each user, we can determine people who are highly correlated to a specific user, in terms of event activity. These correlated users, would then be "trustworthy."

The activity correlation measure $\delta(U_1,U_2)$ between two users U_1 and U_2 is then proportional to the Hausdorff event similarity with the similarity measure ES:

$$\delta(U_1,U_2) = S_H(E_1,E_2 \mid ES). \quad <12>$$

Note that it is important to explicitly compute event annotation similarity – co-participation in itself not enough. We need to further establish *how* people annotate events, including those that are shared.

Now it is straightforward to develop the iterative mechanism to propagate trust in the network. For any given user *k*, the trusted cohorts in the network are computed as follows:

1. We normalize the activity based trust with respect to each of the other members in the network, such that the sum of the trusts adds up to unity.
2. The update equation:

$$\mathbf{t} = \alpha \cdot \mathbf{A} \cdot \mathbf{t} + (1-\alpha) \cdot \mathbf{p}^k, \quad <13>$$

Where the **t** is the trust vector, **A** is the data driven, row-normalized activity correlation matrix, $\mathbf{p}^k$ is the *a priori* trust vector due to the k^{th} user and α is a weighting factor. Each dimension of the trust vector **t** is a real-valued positive number, indicating the degree of trust. The equation states that the trust vector *for each user*, is obtained through iteration over the weighted sum of the activity correlation matrix and the *a priori* user defined trust vector. Note that this is just the familiar page rank equation with bias vector $\mathbf{p}^k$ [5,9]. In <13>, **t** is initialized to **0**.

The trust vector forms a trusted network context – i.e. each user only receives recommendations from the trusted sub-network. We next show how the trust vector can be combined with the similarity and co-occurrence matrixes to determine recommendations.

6 GENERATING RECOMMENDATIONS

We now discuss how to combine similarity and co-occurrence values with trust, to determine event annotations. Let us assume that we are trying to annotate an image for a specific user *k*. Let us further assume that the size of the entire social network is N. Our approach generalizes to a query on arbitrary facet, but we shall restrict ourselves to recommendations when the query is an image, as this is the familiar annotation scenario. Then we proceed as follows:

1. *Trust*: We first compute the annotator trust for each member in the social network with respect to the user *k*. The user will provide the system with *a priori* trust estimates (these are binary values) for some of her friends. Then the system will estimate trust over the entire social network using the iterative procedure stated in equation <13>. This will determine a real

valued number for each member of the social network thus creating the trust vector $\mathbf{t}^k$.

2. *Similarity*: Compute the global similarity matrix $\mathbf{M}_s$ using the procedure outlined in section 4.1.4.
3. *Co-occurrence*: Compute the global co-occurrence matrix $\mathbf{M}_c$ as follows:

$$\mathbf{M}_c|(a,b) = \sum_{i=1}^{N} t^k(i)\mathbf{M}_c^k|(a,b), \qquad <14>$$

Where, $t^k(i)$ is the trust of the i^{th} user with respect to user k. $\mathbf{M}_c|(a,b)$ is derived similar to equation <11> except that the personal co-occurrence matrix of each user is modified by the trust of that user with respect to user k. Note that in this equation the global co-occurrence matrix is computed over the entire network, without thresholding the trust value. If there are scalability concerns in a large network, then the summation can be done over a subset of the network, where each member's trust with the user k exceeds an optimized trust threshold.

Given the trust vector $\boldsymbol{t}^k$, and global similarity ($\mathbf{M}_s$) and global co-occurrence matrix ($\mathbf{M}_c$), we can now determine the recommendations for the query $\mathbf{q}$ as follows:

$$\mathbf{y} = \mathbf{M}_c\mathbf{x} + \mathbf{q},$$
$$\mathbf{x} = \mathbf{M}_s\mathbf{y} + \mathbf{q}, \qquad <15>$$

Where, $\mathbf{q}$ is the query vector, $\mathbf{x}$ and $\mathbf{y}$ represent the similarity and co-occurrence affinity vectors respectively.

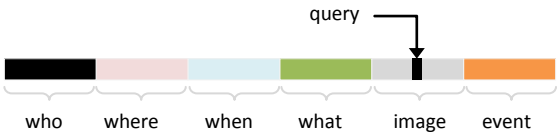

Figure 5: The query vector $\mathbf{q}$ has six facets. The query can be along any of the six facets. Typically, we would query for an image, and the iterative process would retrieve top ranked annotations for the other five facets.

It is useful to review the composition of the query vector $\mathbf{q}$ and vectors $\mathbf{x}$ and $\mathbf{y}$. All the three vectors have six parts – who, where, when, what, image, event (ref. Figure 5). Both vectors $\mathbf{x}$ and $\mathbf{y}$ are initialized to $\mathbf{0}$. The query vector $\mathbf{q}$ typically contains all zeros except for the query dimension, which is represented using 1. At the end of this iterative process, we can determine the recommended annotations per facet by picking the top L annotations per facet (these are the concepts that have the highest score at the end of the iterative process). The iterations are done n times, for reasons of computational efficiency. In our implementation n = 10. At this point, we have recommendations for each event facet using the user's trust vector and the statistical and co-occurrence matrices.

The equation <15> is a variant of the well known HITS algorithm [12]. It suggests that the optimal recommendations are determined by using a coupling strategy between the global similarity matrix, and the trust weighted global co-occurrence matrix. The co-occurrence matrix $\mathbf{M}^i_c$ is weighted by the trust that the user k has with user i. The coupling links the common shared knowledge (similarity between concepts) that exists within the social network with personalized observations (i.e. concept co-occurrences) that the user trusts.

After the user is given the recommendations and has annotated the event with the *who*, *where*, *when* and *what* fields or had added, the system updates the user's similarity and co-occurrence matrices as well as the global similarity and co-occurrence matrixes. The trust vector for this user is updated as the inter-user similarity (ref. equation <12>) will change slightly. Thus, as the users annotate more images, the recommendations will improve as the co-occurrence statistics will become more stable.

7 EXPERIMENTS

We now describe our experimental results. We built an event annotation system that allowed users to create events and then add descriptors to the event (event name, who, where, when what and images). We asked eight graduate students to participate in the experiment. They created 58 events over the course of two weeks, and added 250 images to the collection. Figure 6 shows the event creation page, and the media upload action.

Figure 6: The upload page of our event annotation system. The user can create events, and add event meta-data – text keywords for the facets of *who*, *where*, *when*, *what* and images. Each image also has an author associated with it.

In order to compare the efficacy of the proposed system, we compared it to a baseline SVM based image annotation system. The SVM's were trained using SVMLight [21] and an RBF kernel. In this paper, the query was always an image, though our approach can easily handle queries along each of the other facets, including the case when combinations of facets are specified as queries.

We created two scenarios – global annotation and personal annotation. By global annotation, we imply that the images are annotated by pooling all the images in the social network – similar to what would happen in a Flickr group pool. By personal annotation, we plan to use the annotation framework designed separately per user. For each of the two scenarios, we can compare the SVM based annotation system with the social network based annotation system.

We adopted a modified bagging strategy for each class (in both global and personal cases). Let us assume that we have N positive instances of the class. Then, we constructed five symmetric classifiers, for the same concept. For each such classifier, we picked N negative examples at random from the remainder of the training set, without replacement. Then we obtained average precision and recall for each such classifier and picked the one classifier that maximized the average F-score. We compared this strategy with voting, as well as taking

the average of the five SVM classifier outputs, and this strategy seemed to give slightly better performance.

In the global case, we trained SVM's only for those tags that had more than 10 images associated with them. We found that below this threshold, results were not reliable. This resulted in 31 classifiers (combining classifiers over all facets). Specifically, the classifier breakdown was as follows: who:8, when: 6, where: 10, what: 7. For the social network based annotation, we combined all the co-occurrence matrixes across the social network using uniform trust – this is equivalent to the case that everyone in the network is equally trustworthy annotator.

Figure 7: Photographs for which SVM based global classifiers work well.

For evaluation of the global case, we tested on 50 images, rather than the entire dataset due to computational efficiency reasons. We created a test set of 50 images. However, we tested on one image at a time, thereby having 249 training examples. The computational complexity arises due to the fact that we need to retrain SVM 31 classifiers per test image, using the bagging approach. In our coupling matrix approach with uniform trust, the computational complexity is low – per test image we only need to remove one row and one column for $\mathbf{M}_s$, and adjust the statistics of the co-occurrence matrix $\mathbf{M}_c$.

Table 1 (Global): The table shows that the comparison of SVM with our approach for the global case for 50 images.. H: hits, M: Misses, X: no classifier exists, U: un-decidable. The coupling matrixes (CM) are used with uniform trust over the entire network.

Facets	SVM				CM (uniform)	
	H	M	X	U	H	M
Who	13	23	5	9	22	28
When	11	20	6	13	24	26
Where	12	19	3	16	23	27
What	13	21	8	8	31	19
Event	10	12	22	6	22	28

In addition to the familiar hits (H) and misses (M), we introduce two new testing parameters to make the comparisons to SVM's more nuanced. Since in the global case we can only train 31 classifiers, there will be concepts that cannot be classified at all, due to the small number of samples for that concept. Whenever we encounter such a concept, rather than giving a negative result for SVM, we explicitly acknowledge it under the column X (X: no classifier exists). The other new category is un-decidable (U). This designation implies that *all* of the 31 classifiers give a negative result for this image. Note that traditionally, both categories X and U would be counted as a miss. We classify the output of an SVM or our coupling matrix based approach as a hit, when the concept is a match with one of the top three recommendations of the classifier / coupling matrix. Hence, for each image, we would generate three recommendations per facet and check if a match exists.

The results in Table 1 are interesting. They reveal that for the social network, and for small datasets, the SVM is significantly outperformed by the coupling matrix based image annotation system. These results are for the global case, when we assign uniform trust. Note that in the coupling matrix case, since there are no explicit classifiers, we will never have categories X and U appear. In Figure 7, we see an example, where SVM based classifiers work well.

Figure 8: Photographs for which the coupling matrix based image annotation recommendations work well.

In the personal case, we trained SVM classifiers per person in the network. Now, there are far fewer images per tag. Unfortunately, using the threshold of 10 for the personal case would have left us with just four classifiers over all facets over *all* the members. Instead we decided to train SVM's with five positive examples each. This way we ended up with 28 classifiers. Specifically, the classifier breakdown was as follows: who:9, when: 4, where: 6, what: 9. We note these are totals, *over all users.* Note also that the number of classifiers is less than in the global case (28 vs. 31). This is because when we begin to construct the classifiers per user, there are fewer cases when the number of positive image examples needed to train a concept classifier exceeds five.

Table 2 (Personal): The table shows that the comparison of SVM with our approach for the personal case for 250 images. The coupling matrixes (CM) are used with the trust vector *corresponding to the owner* of the image, over her entire network.

Facets	SVM				CM (network)	
	H	M	X	U	H	M
Who	45	81	62	62	183	67
When	51	96	73	30	167	83
Where	62	76	59	53	179	71
What	72	89	23	66	204	46
Events	0	0	250	0	153	97

For the coupling matrix based approach, we computed the trust vector *for the author of the test image* and used it to compute trust adjusted global co-occurrence matrix (ref. eq. <14>). Then

as before, the system iterated until convergence and then three annotations were provided per facet.

Table 2 shows the classification results aggregated over all the users. It shows that SVM's perform very poorly when compared to the coupling matrix case – this is not surprising for two reasons – the *minimum* number of images used to train an SVM classifier per concept is five. Hence some of the classifiers will not generalize well over the training data.

For both the global and the personal cases, we see that the SVM based approach works poorly compared to the coupling matrix based approach. These are preliminary results, and we currently working to replicate these results on a larger dataset.

8 CONCLUSIONS

In this paper we discussed an event-centric approach to media annotation, which incorporated social network trust. We observed that the tag distribution for a photo pool in Flickr followed the familiar power law distribution found in online social networks. These observations had consequences for concept based annotation systems, in terms of learnability, scalability and semantic variability. Hence we did not develop per-concept classifiers. Instead, our approach to annotation was motivated by web search algorithms such as HITS and PageRank.

We defined events to be a real-world occurrence, which may be described using attributes such as images, and facets such as who, where, when, what. A key idea was that media (including images and text) are event meta-data – i.e. they are descriptions of the event.. Given a social network and events, we showed how to compute event concept similarity, concept co-occurrence and annotator trust. Our image annotation algorithm was a variant of the well known HITS algorithm. The optimal recommendations were determined by using a coupling strategy between the global similarity matrix, and the trust weighted global co-occurrence matrix. The coupling links the common shared knowledge (similarity between concepts) that exists within the social network with personalized observations (i.e. concept co-occurrences) that the user trusts. Our preliminary experimental results when compared to traditional SVM based concept classifiers are promising. We planning to extend this work with experiments on larger datasets as well as incorporating event hierarchies and semantic relations.

9 REFERENCES

[1] *Flickr* http://www.flickr.com.

[2] *SEraja* http://www.seraja.com/.

[3] L. V. AHN and L. DABBISH (2004). *Labeling images with a computer game*, Proceedings of the SIGCHI conference on Human factors in computing systems, 1-58113-702-8, ACM Press, 319-326, Vienna, Austria.

[4] C. ANDERSON (2006). The long tail : why the future of business is selling less of more. Hyperion 1401302378 New York.

[5] S. BRIN and L. PAGE (1998). *The anatomy of a large-scale hypertextual Web search engine.* Computer Networks and ISDN Systems **30**(1--7): 107--117.

[6] E. CHANG, K. GOH, G. SYCHAY and G. WU (2003). *CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines.* IEEE Transactions on Circuits and Systems for Video Technology **13**(1): 26-38.

[7] A. K. DEY (2001). *Understanding and Using Context.* Personal and Ubiquitous Computing Journal **5**(1): 4-7.

[8] P. DOURISH (2004). *What we talk about when we talk about context.* Personal and Ubiquitous Computing **8**(1): 19-30.

[9] Z. GYONGYI, H. GARCIA-MOLINA and J. PEDERSEN (2004). *Combating web spam with TrustRank*, Proceedings of the 30th International Conference on Very Large Data Bases (VLDB) 2004, Toronto, Canada.

[10] D. HUTTENLOCHER, G. KLANDERMAN and W. RUCKLIDGE (1993). *Comparing Images Using the Hausdorff Distance.* IEEETPAMI: IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(9): pp. 850-863.

[11] L. KENNEDY, A. HAUPTMANN, M. NAPHADE, A. HAUPTMANN, J. R. SMITH and S.-F. CHANG (2006). *LSCOM Lexicon Definitions and Annotations Version 1.0*, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, ADVENT Technical Report #217-2006-3, March 2006., Columbia University.

[12] J. KLEINBERG (1999). *Authoritative sources in a hyperlinked environment.* Journal of the ACM **46**: 604-632.

[13] B. LI, K. GOH and E. CHANG (2003.). *Confidence-based Dynamic Ensemble for Image Annotation and Semantics Discovery.*, ACM International Conference on Multimedia., 195-206, Berkeley, CA.

[14] H. LIU and P. SINGH (2004). *ConceptNet: a practical commonsense reasoning toolkit.* BT Technology Journal **22**(4): pp. 211-226.

[15] G. A. MILLER, R.BECKWITH and C.FELLBAUM (1993). *Introduction to WordNet : An on-Line Lexical Database.* International Journal of Lexicography **3**(4): 235-244.

[16] M. NAAMAN, H. GARCIA-MOLINA, A. PAEPCKE and R. B. YEH (2005). *Leveraging Context to Resolve Identity in Photo Albums*, Proc. of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2005), June 2005, Denver, CO.

[17] B. SHEVADE, H. SUNDARAM and M.-Y. KAN (2005). *A Collaborative Annotation Framework*, Proc. International Conference on Multimedia and Expo 2005, Jan. 2005, Amsterdam, The Netherlands.

[18] B. SHEVADE, H. SUNDARAM and L. XIE (2007). *Modeling Personal and Social Network Context for Event Annotation in Images*, Proc. Joint Conf. on Digital Libraries 2007, Jun. 2007, Vancouver, Canada.

[19] B. STERLING (2005). *Order Out of Chaos*. Wired. **13.04** http://www.wired.com/wired/archive/13.04/view.html?pg=4.

[20] S. H. STROGATZ (2001). *Exploring complex networks.* **410**(6825): 268.

[21] SVMLIGHT *http://svmlight.joachims.org/*

[22] M. TRURAN, J. GOULDING and H. ASHMAN (2005). *Co-active intelligence for image retrieval*, Proceedings of the 13th annual ACM international conference on Multimedia, 1-59593-044-2, ACM Press, 547-550, Hilton, Singapore.

[23] T. V. WAL (2006) *Off the Top: Folksonomy* http://www.vanderwal.net/random/category.php?cat=153.

[24] U. WESTERMANN and R. JAIN (2007). *Toward a Common Event Model for Multimedia Applications.* IEEE Multimedia **14**(1): 19-29.

[25] A. WILHELM, Y. TAKHTEYEV, R. SARVAS, N. V. HOUSE and M. DAVIS (2004). *Photo Annotation on a Camera Phone*, ACM Conference on Human Computer Interaction, Apr. 2004, Vienna, Austria.