A COLLABORATIVE ANNOTATION FRAMEWORK

Bageshree Shevade *Hari Sundaram* *Min Yen-Kan*
Arts, Media and Engineering, ASU National University of Singapore
Email:{bageshree.shevade, hari.sundaram}@asu.edu, kanmy@comp.nus.edu.sg

ABSTRACT

This paper describes our system that enables members of a social network to collaboratively annotate a shared media collection. The problem is important since online social networks are emerging as conduits for exchange of everyday experiences. Our collaborative annotation system provides personalized recommendations to each user, based on (a) media features, (b) context, (c) commonsensical relationships and (d) linguistic relationships. We also develop novel concept specificity and abstractness / concreteness measures that further adapt the recommendations to the specific concept. Our preliminary user studies indicate that the system performs well and is more useful as compared to standard web browser recommendation schemes.

1. INTRODUCTION

In this paper, we develop a novel system that allows a network of friends to collaboratively annotate shared media. This problem is important in several contexts: (a) people share and archive events associated with everyday experiences due to easy availability of digital cameras, and (b) networked exploration frameworks that allow people in a social network to exchange ordinary everyday experiences are predicated on the presence of annotation.

There has been prior work in creating collaborative annotation systems [4,6,8]. In [8], the authors explore a collaborative annotation system for mobile devices. There they used appearance based recommendations to suggest annotations to mobile users. In [6], the authors describe a collaborative annotation procedure for scientific visualization tasks, that can be done remotely. In [4], the authors study how annotations undergo transitions when they move from a personal to a shared environment. A key innovation in our approach is to augment the feature based recommendation systems with a common sense toolkit and linguistic relationships, thus making the recommendations more personalized and useful.

In our approach, our collaborative recommendation system consists of the following components: (a) media and its features, (b) user / group context, (c) common sense relationships and (d) linguistic relationships. The user annotates the images using a web-based interface. As the user begins to annotate images, the system provides personalized recommendations using a combination of low-level, common-sense and linguistic features. It also provides group recommendations based only on low-level features. A key innovation here is a measure of abstractness / concreteness and concept specificity, that allows *us to adaptively change the number of recommendations based on the specific concept.*

Once the user has finished annotating an image, the system creates positive example image sets (or clusters) for the associated annotation words within each field (*who*, *when*, *where, what*). The clusters are based on annotation words/concepts entered by the users and not on automatic grouping of low-level features. These clusters will help the annotation process improve *for all users of the network.*

The rest of this paper is as follows. In the next section, we formulate our problem statement and present the solution. In Section 3, we describe the components of our collaborative annotation system. Section 4, describes our collaborative annotation algorithm in detail. In Section 5, we present our experimental results. Finally, we present our conclusions in Section 6.

2. PROBLEM STATEMENT

Our goal is to develop a system that enables a network of users to collaboratively author shared media. Since annotation is crucial to networked exploration frameworks, we need to do the following:

- Provide tools that will facilitate minimal authoring of shared media by providing recommendations for annotation.
- Devise methods that will recommend using low-level media features so that it exploits the fact that members of a social network, share activities and events and hence recognize shared objects.
- Personalize the recommendations using context and linguistic features as well as commonsensical relationships.

3. SYSTEM COMPONENTS

Our collaborative annotation system consists of the following components – (a) media and its features (b) context (c) commonsensical relationships and (d) linguistic relationships. We now discuss each of these in detail.

3.1 Features

In our system the media consists of images associated with everyday activities. The feature vector for images comprises color, texture and edge histograms. The color histogram consists of 166 bins in HSV space. The system extracts Tamura texture [2] from images. The texture histogram consists of 3 bins corresponding to contrast, coarseness and directionality of the image. The edge histogram [2] consists of 71 bins that incorporates curvature and edge directionality. We then concatenate the three histograms to get a final composite histogram of 240 bins. The low level feature distance between two images *i* and *j* is then given as:

$$d(i,j) = \sqrt{\sum_{k=1}^{N} (h_i^k - h_j^k)^2}, \qquad <1>$$

where N is the total number of bins, and h_i^k and h_j^k are the corresponding bins of images *i* and *j*.

3.2 Context

User context models are crucial to collaborative annotation systems as they help in giving personalized recommendations to

0-7803-9332-5/05/$20.00 ©2005 IEEE

each user. The dictionary definition of context is given as: *the interrelated conditions in which something exists or occurs.* These conditions could be the physical location, time, user's activity and past actions, environment etc. [7].

In our system, the user's context model comprises of (a) the initial static user profile which includes demographic information like age, background, hobbies/interests etc. (b) statistical information like number of images contributed in the shared social network and (c) usage statistics which includes the words she has used for annotation and their frequency. Frequency count is maintained for each of the *who*, *where*, and *what* fields of an image. Modeling the user-context using frequency count is intuitively useful and reliable as the shared media consists of everyday events and these events recur.

In our system, the group context model comprises (a) the images uploaded by all the members of the group and (b) the annotation words used by all the members of the group. Our system uses the group context to provide each user, recommendations for the group list using low-level features.

3.3 Commonsensical and linguistic relationships

In our system, semantics are incorporated through the use of ConceptNet [3]. ConceptNet is a large repository of common sense knowledge and is suitable for making practical inferences over text. The repository supports twenty semantic relationships like *"capableOf"*, *"locationOf"*, *"usedFor"* etc. Since the media consists of everyday events, we believe that the use of ConceptNet, will enhance the quality of recommendations. ConceptNet links the group recommendations that are based on low-level features with the concepts in the user profile, using assertions about everyday events and activities. We measure the average distance between the group recommendation and the user profile using measures in [1]. Then, if the distance is less than a threshold, we shall use ConceptNet to determine the context of the concept [3].

In our system, linguistic relationships are incorporated through the use of WordNet [5]. WordNet is an online lexical database, created by linguists that specifies semantic relationship between concepts. WordNet organizes English nouns, verbs and adjectives into synonym sets called *synsets* which represent one unique lexical concept. Each synset also contains multiple words or word forms that are synonyms of each other.

3.4 Concept Specificity and Abstract / Concrete Measures

In our framework, we use the hypernym / hyponym relationship supported by WordNet to determine if a concept is *abstract* or *concrete*. Concrete concepts are those which can be sensed using the five senses. This is useful, since we conjecture that abstract concepts (love, anger etc.) are more likely to be interpretive and individualistic as opposed to concrete concepts (water, ball etc.), whose meaning is likely to be shared within the social network.

The system computes a measure of *abstractness / concreteness* and *specificity* for the concept and uses it to determine the number of filtered concepts to return as recommendations. The system returns a larger number of filtered recommendations when a concept is abstract than when it is concrete.

WordNet organizes all its noun synsets into hierarchies that are headed by a synset called a *unique beginner*. Some of these unique beginner synsets are *"entity, physical thing"*, *"abstraction"*, *"state"*, *"event"* etc. We have classified these beginner synsets into two broad classes *"abstract"* and *"concrete"* based on standard linguistic references. Thus, if a noun synset terminates in a beginner synset which is classified as abstract, then it is considered abstract, otherwise it is considered as concrete. We have also classified verb synsets into *"abstract"* and *"concrete"* classes to determine if the verb form of a concept is abstract or concrete. We then combine measures of both noun and verb forms to get the final abstractness/concreteness and specificity measure.

Given a concept word *w*, we first extract all the noun synsets for that word. For every noun synset, we then determine the root nodes and maintain the hop count to the root nodes. This hop count is then averaged over the number of paths existing to the root nodes. This is the up-distance U_d and is given as:

$$U_d = \frac{\sum_{j=1}^{M} h_j}{M}, \qquad <2>$$

where M is the total number of paths to the root nodes and h_j is the number of hops. The system then determines all the leaf nodes of the noun synset. The system again maintains a different hop count and averages it over all the paths existing to the leaf nodes. This is the down-distance D_d and is given as:

$$D_d = \frac{\sum_{k=1}^{N} h_k}{N}, \qquad <3>$$

where N is the total number of paths to the leaf nodes and h_k is the hop distance. The system now computes a specificity measure S^n_i for a noun synset *i* as:

$$S_i^N = \frac{U_d}{(U_d + D_d)}, \qquad <4>$$

where U_d is the up-distance and D_d is the down-distance. A value of 0 indicates that the synset is very general and a value of 1 indicates that it is very specific. This is intuitive since a synset which is close to the leaf will have a large U_d of "is-a" relationships and hence will be very specific and a synset close to the root will have a large D_d and will be very general.

The system also determines the synset probability using tag count; which is the frequency of usage of that synset. The frequency value is normalized over all the noun synsets of concept word *w*. The noun synset probability NP_i is then given as:

$$P_i^N = \frac{f_i}{\sum_{i=1}^{K} f_i}, \qquad <5>$$

where K is the total number of noun synsets and f_i is the tag count of synset *i*. The system then computes the final noun specificity measure, S^N_{final}, for the concept word *w* as:

$$S_{final}^N = \sum_{i=1}^{K} P_i^N \bullet S_i^N, \qquad <6>$$

where K is the total number of noun synsets. The system also determines the noun abstract / concrete property measure for the concept word *w* as:

$$A^N = \sum_{i=1}^{K_c} P_i^N - \sum_{j=1}^{K_a} P_j^N, \qquad <7>$$

where K_c is the total number of concrete noun synsets and K_a is the total number of abstract noun synsets and P^N_i is the noun synset probability of synset *i*.

The system then uses the same procedure to compute the verb specificity measure and the verb abstract / concrete property using all the verb synsets for the concept word *w*. The one difference is that WordNet does not have a root synset for verbs – instead, each verb synset is categorized, and the authors label the categories as concrete or abstract. Then all the equations derived for nouns, hold for verbs as well. The final specificity measure S_{final} and final abstract / concrete property measure A, for concept word *w* are then given as:

$$S_{final} = \alpha S^N_{final} + \beta S^V_{final}, \quad \alpha = \frac{T^N}{T^N + T^V}, \quad \beta = \frac{T^V}{T^N + T^V}, \qquad <8>$$
$$A = \alpha A^N + \beta A^V,$$

where T^N is the total number of noun synsets and T^V is the total number of verb synsets, and $\alpha+\beta=1$, where the superscript indicates the noun / verb measures. We have just described a framework to adapt the recommendations of the group, for each user, based on linguistic and commonsensical relationships.

4. COLLABORATIVE ANNOTATION ALGORITHM

In this section, we shall discuss the algorithm for the collaborative annotation system in detail. The goal is to provide recommendations as the user is trying to annotate images uploaded by her. Let us assume that the user wishes to annotate an image *a* with the *who*, *where*, *when*, and *what* fields. Let us also assume that the database contains N clusters for annotations within each field.

As the user annotates images, the system creates positive example image sets for the associated annotations. The system forms clusters for each distinct annotation introduced in the system. Note that these clusters are *not* created using clustering techniques such as k-means, but are due to the annotation groupings. The system is shown in Figure 1. For an un-annotated image, the system provides two kinds of recommendation lists: (a) personal and (b) group.

4.1 Feature based Group Recommendation

The group recommendation for each field is obtained by computing the low-level feature distance between image *a* and the N cluster centers. The system then presents the top three closest cluster center words as recommendations in the group list. The images that comprise these clusters have been annotated by the other members of the social network with annotation. So, as the users introduce new annotation words in the system, new clusters get created corresponding to those words and the images become positive examples of those clusters.

4.2 Concept Filtering

The system filters the group recommendation list by the user profile to get additional personal recommendations for the *what* field of the image. This is done by computing the semantic distance using ConceptNet, between every concept in the user's profile and the concepts returned in the group recommendation list. When the semantic distance is less than an optimized threshold, the system uses the ConceptNet toolkit to get a list of concepts which are in the context of user profile concept but biased by group recommendation concept [1].

In order to determine the number of filtered concepts to return as recommendations, the system computes a measure of abstractness/concreteness and specificity for the group recommendation concept that matches the user profile. This is done using equation <8>. This measure is used to vary the

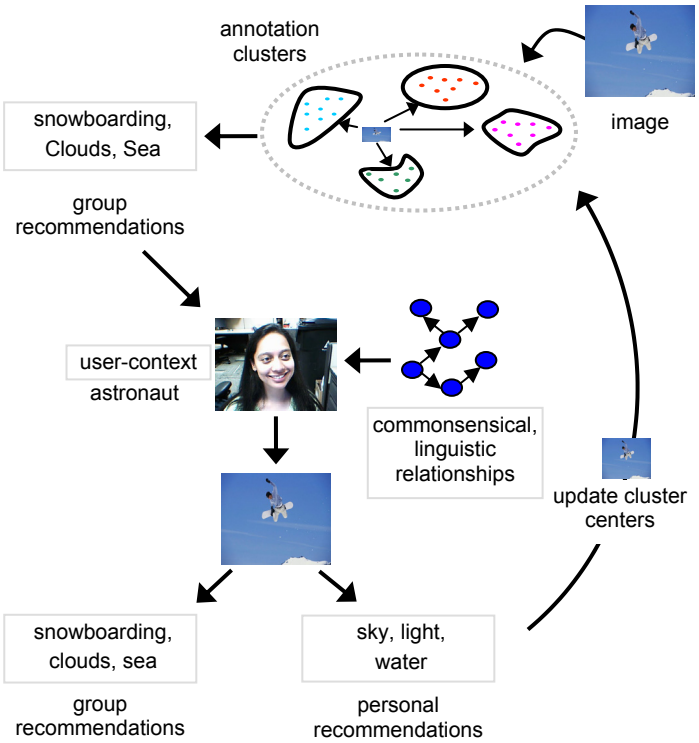

Figure 1: The personal and group recommendations are generated using low-level features, user-context, group-context, common-sensical and linguistic relationships.

degree of personalization by varying the number of filtered concepts returned. We map the two dimensions of abstract/concrete property and specificity on a scale of 1 to 16 such that the number of filtered concepts returned, conform to the following order:

$$N_{ag} > N_{as} > N_{cg} > N_{cs}, \qquad <9>$$

where N_{ag} is the number of filtered concepts returned for a concept that is *abstract* and *general*, N_{as} is for a concept that is *abstract* and *specific*, N_{cg} is for a concept that is *concrete* and *general* and N_{cs} is for concept that is *concrete* and *specific*.

4.3 Frequency based Personal Recommendation

The personal recommendation list is obtained from the frequency count of the annotation words used by the user. As the user annotates images, the system maintains a frequency count within each field for each annotation word used. The system then picks the three most frequently used words within each field to generate the personal list.

4.4 Updating the System

When the user has annotated image *a* with the recommendations provided or by entering her own annotations, the system treats image *a* as a positive example of the all the annotations associated with it. The system thus creates semantic clusters

corresponding to all annotation words that exist in the system. If the user has introduced a new annotation word in the system, then system creates a new cluster for the annotation with only image *a* as the positive example. The system also updates the user profile with the words that the user has chosen for annotation, thus making the user profile dynamic.

5. EXPERIMENTS

We conducted three preliminary experiments to evaluate the quality of recommendations provided and to measure the utility of the adaptive recommendation list.

In order to evaluate the annotation system, members of the network which consisted of four graduate students at ASU, were asked to upload and annotate shared media using this system. The system was also seeded with an initial user profile of all members. As the users annotated images, the system maintained a count of the recommendations that were chosen by the user to annotate her media. We chose to evaluate the system against a baseline recommendation system, commonly found in web browsers – recommendations were given on recently used (RU) annotations. Users were presented with the two systems and they annotated around 30 images from everyday activities, in each system.

As the results in Table 1 indicate, our collaborative annotation system performed better than the web browser systems. Since the collaborative annotation system was based on ConceptNet and WordNet, there was an increase in the number of recommendations provided. This is intuitive, since the media consisted of everyday events of members in a social network, and so we expect to see similar people, places and activities across images. As a result, users could choose more annotations from the recommendation list and add fewer new annotations, thus reducing the time spent in annotating media. However, the difference between the two systems is not very large, since the images belonged to very few (four) events. On the average, there were four images per person. So, when images span a large number of different events, then RU will not be very useful.

Table 1: Comparison of our collaborative annotation system against recommendation schemes in web browsers. Our system performed better as users picked more recommendations and added fewer new annotations as compared to web browser systems.

	Collaborative Annotation System	**Web browser system (RU scheme)**
No. of new annotation words added	**36 / 110**	40 / 106
No. of recommendations chosen	**74 / 110**	66 / 106

In order to determine the utility of the adaptive recommendation list, the system kept track of the all the images when the user did not choose from the recommendation list but added her own annotation words. The system then determined if the annotation word introduced by the user or its synonyms were encompassed by a larger returned recommendation list. The results show that only 9% of the words belonged to the extended list. This indicates that the utility of the system is good as it reduces the time to annotate by not giving a larger list.

The system also measured the difference in utility and quality of recommendations, if only the most commonly used sense of a concept word was used to determine the specificity and abstract / concrete property as opposed to using all the synsets. The average difference in the size of the adaptive list was 0.5 when measured over 84 concepts. Since the difference is very small, we plan to reduce computational complexity by using the most common synset.

6. CONCLUSIONS

In this paper, we presented a novel collaborative annotation system that enables members of a social network to annotate shared media. The system provides recommendations based on (a) low-level features, (b) context, (c) commonsensical and (d) linguistic relationships. For each un-annotated image, the system uses low-level features to make initial recommendations. These are personalized using an adaptive framework that utilizes the user context as well as commonsensical and linguistic relations. We conducted preliminary experiments. They indicate that our collaborative annotation system performed well. In the future, we plan to use sophisticated classification techniques to improve the feature based recommendations. We also plan to address scalability and performance issues.

7. REFERENCES

[1] P. APPAN, B. SHEVADE, H. SUNDARAM, et al. (2005). *Interfaces for networked media exploration and collaborative annotation*, to appear in International conference on intelligent user interfaces 2005, also AME-TR-2004-11, Jan. 2005, San Diego, CA.

[2] A. K. JAIN (1989). Fundamentals of digital image processing. Prentice Hall Englewood Cliffs, NJ.

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

[4] C. C. MARSHALL and A. J. B. BRUSH (CHI, 2002). *From personal to shared annotations.*

[5] G. A. MILLER, R. BECKWITH, C. FELLBAUM, et al. (1993). *Introduction to WordNet: An On-line Lexical Database.* International Journal of Lexicography **3**(4): 235-244.

[6] D. P. G. SEAN E. ELLIS (2004). *A Collaborative Annotation system for data visualization.* Proc. of the working conference on Advanced visual interfaces.

[7] H. SRIDHARAN, H. SUNDARAM and T. RIKAKIS (2003). *Context, memory and Hyper-mediation in Experiential Systems*, 1st ACM Workshop on Experiential Telepresence, in conjunction with ACM Multimedia 2003,, AME-TR-2003-02, Nov. 2003., Berkeley CA.

[8] A. WILHELM, Y. TAKHTEYEV, R. SARVAS, et al. (CHI, 2004). *Photo Annotation on a Camera Phone.*