Exploiting Tags for Concept Extraction and Information Integration

Martha L. Escobar-Molano
SET Corporation
1005 Glebe Road, Suite 400
Arlington VA

Antonio Badia
Computer Engineering and Computer Science
University of Louisville
Louisville KY

Rafael Alonso
SET Corporation
1005 Glebe Road, Suite 400
Arlington VA

Abstract—**The use of tags to annotate content creates an opportunity to explore alternatives to automate the process of extracting semantics from data sources. Semantic information is needed for many complex tasks like Concept Extraction and Information Integration. In order to establish the value of user-generated annotation, this paper presents two experiments on which only user tags are used as input. At the core of semantic extraction is the identification of concepts and relationships that are present in the data. We show, through an experimental study on tagged photographs, how to extract concepts associated with photographs and their relationships. Our experiments demonstrate that supervised machine learning techniques can be used to extract a concept associated with a photograph with an overall precision score of 80%. Our experiments also show that a variation of the Jaccard similarity coefficient on sets of tags can be used to determine equivalence relationships between the concepts associated with these sets.**

I. Introduction

Collaborative tagging has become very popular on the web. The core idea is that users annotate content with free-form *tags*. These tags can be any string of characters that the user deem useful to describe content. Because annotations come from many users with many points of view, there is great interest in harnessing this *wisdom of the crowds* to leverage more semantics.

However, the value of tags (and collaborative efforts in general) has been debated. Tags are created by users without any centralized control, and therefore there is no guarantee that the joint effort will *converge*, that is, create a consistent and representative picture of the data being tagged. Also, tags come from *uncontrolled vocabularies*, hence users are free to use whatever they want as tags. Therefore, tags might have problems such as spelling errors and ambiguity.

Collaboration without explicit control has been contrasted with more hierarchical, organized approaches. In particular, tagging and the resulting *folksonomies* have been compared with thesauri and ontologies. Ontology creation is a highly edited, tightly controlled process carried out by a small team (or even a single person) of experts. As a consequence, the result is supposed to be a high quality representation. Using folksonomies is a *bottom up* effort, the product of a decentralized body, while ontologies are *top down* efforts.

Both approaches have disadvantages, and several authors have pointed out the potential of collaboration.. Ontologies are very resource-intensive to create, are often brittle and subjective. The use of tags to annotate web content created an opportunity to explore alternatives to automate the process of extracting semantics from data sources. Several studies have shown the semantic value of folksonomies [1], [2], [3]. Compared to ontologies, the use of folksonomies to extract semantics has the advantage of representing the collective intelligence instead of the perception of few experts [4].

One way to test the semantic value of collaborative efforts and, in particular, tagging is to use (exclusively) tags for tasks that would ideally require semantic information about a domain. Two such tasks are *Concept Extraction* and *Information Integration*. In Concept Extraction, we are given a set of concepts, a set of data instances, and we must determine how the given concepts classify the given data by assigning one (sometimes more) concept to each instance. In Information Integration, we are given two sets, each one including some already classified instances (and the concepts they fall under), and we are asked to merge the two sets into a single one so that in the resulting set similar concepts (and their associate instances) are represented by a single concept, while different concepts are kept disjoint. Both problems are long standing challenges in Artificial Intelligence and Databases, still unsolved in their entirety in spite of many years of research and some advances. There is consensus that, for both problems, to be solved to a satisfactory degree semantic information about the data is needed. Ontologies have been used to represent semantics understandable by computers. Thus, a widely used approach is to have knowledge engineers build and maintain ontologies for the underlying domains and define algorithms that take advantage of this added knowledge.

In this paper, we use tags to carry out an analysis of several data sets. We focus on data analysis for Concept Extraction and Information Integration and rely *exclusively* in user tags. Our goal is to find out if sets of user tags carry significant *semantic* information about the items that they tag. We measure semantic content indirectly, through several activities which are usually conceived as reflecting the semantics of the data.

We obtained data from several social photography sites, sites where photographers offer their pictures for sale, and users can

This material is based upon work supported by the U.S. Army Aviation and Missile Command under Contract No.W31P4Q-06-C-0473

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330

see and tag pictures on the site freely. Because pictures must be added to the site under a certain category, chosen from a predetermined taxonomy, each picture comes with both a set of tags and a conceptual tag.

At the core of information integration is the description of concepts (or types of objects) and relationships that exist in the data sources to integrate. Our experimental results on tagged photographs show that concepts can be extracted from tags and tags can be used to determine relationships between concepts. We used supervised machine learning techniques on tagged photographs to map a set of tags into concepts. In another set of experiments, we attempt to determine relationships between concepts in different sites based on their tag set. We assume that each Web site has a certain audience, and that audiences from different Web sites are (mostly) disjoint. In spite of being built by different communities, the tag sets showed enough consistency across sites to help differentiate between pairs of similar and dissimilar categories. Overall, our results show that, in spite of several problems, tags do carry significant semantic information and can be meaningfully exploited for both tasks.

This paper is organized as follows. In section II we describe some relevant previous work. In section III we describe some characteristics of tag sets that we have empirically determined. In section IV we describe our approach in general terms and fix some vocabulary. Then in section V we describe in detail the experiments we carried out using exclusively tags for data analysis, including a discussion of results. Finally, we close in section VI with some preliminary conclusions and a discussion of further work.

II. Related Work

Collaborative tagging has become very popular in the web. Researchers have been studying the behavior of tags in social bookmarking systems such as `del.icio.us`. Descriptions of URLs represented by tags attached by users and their frequency were found to reach a stable pattern in which the frequency of each tag is nearly a fixed proportion of the total frequency of all tags used [5]. Tags were found to occur as text in over 50% of the pages they annotate and 20% of the tags do not occur in the page text they annotate nor the back link text nor the forward link page text [6].

Tags in social Web sites such as `del.icio.us` have been found to be good summaries of the corresponding web pages [7], [8]. There have been several studies on semantics extraction from tags in social web sites. Associations between users, tags, and annotated objects are represented by a tripartite hypergraph and network analysis tools have been used to cluster tags in `del.icio.us` [9] and to derive emergent semantics in [3]. In this last paper, the users, tags and data involved are mapped into a "conceptual space". This conceptual space is not given, but constructed as a probability distribution over a vector space with a fixed number of dimensions. The conceptual space is assumed to generate the set of (user, tag, item) observed in the data, and the Expectation-Maximization algorithm is used to reconstruct it. While the framework is similar to ours, here we assume that the conceptual space is given and not generated, and concentrate on other issues: our work on concept extraction and information integration does not require us to assume any latent variables, unlike the approach to clustering and topic distillation in [3]. In [7], the data set of (user, tag, item) is represented as a bipartite graph of tags and item, with the links connecting them reflecting user count. From the resulting tag-by-item matrix (TI), two association matrices of tag-by-tag (TT) and item-by-item (II) are built, using an iterative algorithm similar to Page Rank: at each step, TT is updated by "multiplying" TI and II (for a given pair of tags t_1 and t_2, their vectors in TI are compared item by item, using II to determine item similarity); similarly, II is updated by "multiplying" the inverse of TI and TT. This work assumes, like we do, that similar tags are applied to similar items, and that similar items are characterized by similar tags. Our tasks, however, are different: instead of defining similarity (among tags or items) based purely on the data, we use pre-existing taxonomies to define when data falls under the same concept, and to define when concepts are (dis)imilar, based on the tags.

Extraction of Broader/Narrower relationships between tags have been proposed based on containment relationships between the objects annotated by the tags: Tag A is broader than tag B, if objects tagged by A is a superset of objects tagged by B [9]. However, in practice this would result in a very sparse lattice. The author proposed a relaxation from the containment relationship to overlapping; however the experimental results are inconclusive. Probabilistic models have been used to find associations between tags and unnamed concepts [2]. Users' interests have been added to these models and the resulting models have been applied to search for sources similar to a given source [10]. A hierarchical clustering model was proposed to extract a binary tree of unnamed concepts [2]. However, limiting the hierarchy to a binary tree is unrealistic. Unlike these studies, our work associates named concepts with tags

Recommending tags for images have been studied before [11]. Their approach uses Kernel Canonical Correlation Analysis to learn the correlation between visual content and tags and then use this correlation and tag popularity to recommend tags. While this study expands the tags associated with an image, our work maps set of tags into concepts. The use of tags to improve image search was studied by Kato et al [12]. Their experimental study showed that image search works better for concrete terms (e.g., apple) than for abstract terms (e.g., spring, happy). They focus on improving search when query terms denote abstract concepts. Their study proposes to replace abstract query terms with sets of concrete terms that co-occur with abstract terms as tags in social Web sites that annotate images. While their work expands an abstract term with co-occurring tags, our work maps a set of tags into concepts.

To the best of the authors' knowledge, tags have not been used for the tasks proposed here. There is a very large body of literature on Information Integration, with many different

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330

approaches based on examining different characteristics of data sets (see, for instance, [13]), but the use of user tags has not played a major role in this line of research so far.

III. Tag Sets

In the following, a *page* is simply a Web page. This is due to the fact that in all our experiments we use Web sites that organize photographs, and all sites put each photograph, with all related information (including tags), in a separate Web page. A *category* is a concept in a hierarchy. In our experiments, each Web site has a hierarchy, which is defined by the site owner, and therefore fixes a vocabulary for the whole site. Hierarchies vary from site to site, though. We work with sets of photographs. These photographs and their tags were obtained from the social websites `www.shutterpoint.com` and `www.featurepics.com`. Both Web sites are set up to post and buy photographs. Photographers tag their photos with free-form keywords. They also assign up to three categories to each photograph (in `Shutterpoint`) and a single category in `FeaturePics`). These categories include concepts such as: 'still life', 'general', 'nature', and 'architecture'. Unlike tags, these categories come from a constrained vocabulary. Photographers are presented with a fixed list of categories to choose from. We note that the categories are different on each site, although there is a large amount of overlap. The categories are not organized in a hierarchy in *Shutterpoint*, while in *FeaturePics* categories are organized taxonomically. This taxonomy has maximum depth 5, although most items are at level 3 or less.

Users annotate content with free-form tags. These tags can be any string of characters that the user deem useful to describe content. As a result, these tags have problems, including: spelling errors, polysemy, synonymy, and basic level variation [14]. Polysemy refers to a tag having many senses. Synonymy refers to multiple tags having the same or similar meaning. Basic level of variation refers to users tagging content at different levels of specificity. For example, tags 'car', 'wheeled vehicle', and 'vehicle' describe an object at different level of specificity. Finally, a phenomenon that is particular to tag systems is what we call the *complex tag* issue. Some systems require tags to be single terms (that is, words in the vocabulary of a natural language), while some allow complex terms (like compound nouns, or grammatical constructs like *adjective + noun*). This creates an interesting situation, in that sometimes users want to express complex concepts that require complex terms. When the system does not allow that, users usually resort to creating a *syntactically single term* (that is, one that the system will take as being a single string) by concatenating words with a separator different from the whitespace (for instance, underscore '_' or hyper '-' are commonly used, but capitalization is also utilized). This complicates working with tags considerably, as one must determine exactly how to deal with such complex tags. If it is decided to separate them into simple components, the task can become quite tricky if no obvious separator is used (for instance, when capitalization is used). Also, the resulting terms may not be very meaningful when considered in isolation (for instance, when splitting "blue-eyed-baby"). If it is decided to leave complex tags as such, one must decide if they are tags on their own. At issue is the fact that many times such tags can actually be understood as similar to other, existing tags, perhaps with a different level of specifity (in many cases, the simple tag is literally contained in the complex one; for instance, when tags "baby" and "small baby" appear together). Are these to be considered as different tags or as the same one?

There have been questions as to whether tags sets are usable for data analysis, due to the fact that tag sets have a considerable level of noise, because of lack of editing. In spite of this, Huberman found, in his analysis of a set of tags from `del.icio.us` [5] that tags tend to create *stable* sets over time. He hypothesized that tagging is a social activity, in that the tags chosen by a particular user u were influenced by tags posted by other users before u (tags which u could see). As a consequence, the tag sets tend to *coalesce* over time. He also indicated that tag sets seem to follow a Zipf's law.

In our experiments, described next, we have found that

- tag sets are indeed dirty. Problems we have seen include: singular/plural forms, typos and misspelling, and the complex tag issue.
- tag sets do seem to follow a Zipf law, for a given set of related data items.

To deal with some of these problems, in our work we have treated tags pretty much as keywords in Information Retrieval. We have used case normalization, stemming, dictionary check, and in some cases also synonym check (with WordNet).

One issue that needs to be determined is how much data (or, more precisely, how many tags) constitute a representative sample from which conclusions can be drawn. This involves the practical problem of how much data (or how many tags) must be gathered for experiments. We carried out two experiments to determine the *growth* of tag sets as more data is analyzed. The first experiment gathered a number of pages from within a category in `www.shutterpoint.com`, counted the number of tags thus obtained (before any processing), processed the tags to remove duplicates, plurals and (to a certain degree) misspellings, and then counted the number of *unique, clean* tags left. The results are shown in Table I. A similar experiment was carried out in a different category on the same Web site. However, this time each data set includes the previous one, i.e. each dataset is a superset of previous ones, while in the previous experiment each data set was different (although some overlap was possible, as pages were fetched at random). Note also that in the previous experiment we grow the dataset linearly, and in this second one we grow by doubling the number of pages fetched at each step. The results are show in Table II and, in spite of the differences in setup, are similar to those of the previous experiment: they show two clear tendencies: one, the number of *clean* tags gathered grows less than linearly with the amount of data analyzed; and two, the number of clean tags remains a small percentage of the number of dirty tags, regardless of amount of data captured. The results could be explained by

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330

the expectation that, as more data (and tags) are captured, the likelihood of repetitions increases, and so the number of clean tags would slow its growth.

Nr. of pages	Nr. of *dirty* tags	Nr. of *clean* tags
20	6262	1666
40	12293	2941
60	19334	3895
80	25427	4705
100	30973	5377
120	36272	5942
140	42061	6572
160	45896	6930

TABLE I
GROWTH OF NUMBER TAGS WITH DATA (SHUTTERPOINT'S CHILDREN CATEGORY)

Nr. of pages	Nr. of *dirty* tags	Nr. of *clean* tags
25	6434	2253
50	12639	3561
100	25507	5788
200	51868	9656
400	107577	15692
800	212076	23776

TABLE II
GROWTH OF NUMBER TAGS WITH DATA (SHUTTERPOINT'S TRANSPORTATION CATEGORY)

IV. FRAMEWORK

The typical setup of most past studies has organized data into triples

$$(itemid, tagid, userid)$$

where $itemid$ is an identifier for a data item, $tagid$ is an identifier for a tag (as described above), and $userid$ is an identifier for the user. The triple is intended to express the fact that user $userid$ tagged the item $itemid$ with tag $tagid$. In some cases, some additional *dimensions* are added to this basic fact, *time* being the most common one (i.e. the time at which the user tagging action occurred). In our studies, we disregard the user component purposefully. Our aim is to analyze the result of tagging as a social system, by leveraging the overall result of a group of users interacting with a given data item. Thus, we look at pairs of the form

$$(itemid, tagid)$$

This relation is assumed to be *many-to-many*, that is, an item can be associated with several tags and a tag may be associated with several items.

At the same time, we assume that all items are classified in a family of concepts. Hence, we also assume a binary relation

$$(itemid, conceptid)$$

where *conceptid* is the identifier of a *concept*. Such a relation is assumed to be *many-to-many*, that is, several items are classified under the same concept and several concepts might be assigned to each item. For us, a concept is simply a label which is used to categorize data items; we will not enter to discuss its meaning. Note that usually a hierarchy of concepts, that is, a *taxonomy*, is assumed. However, we will stop at the "bottom" level of such a taxonomy and focus only on the classification of concepts for two reasons:

- in the data sets used in our experiments, most hierarchies (actually all of them with one exception) were 2-level hierarchies, with the top level being an artificial "root" node -that is, the hierarchies really correspond to classifications, and are captured by our conceptualization.
- extracting true hierarchies (that is, not only classifications, but also subsumption relations among given or obtained concepts) is currently an active area of research, but there is no consensus (even no proven method) to carry out such an extraction from tags. Results so far are inconclusive and more research is needed.

.

The above two (binary) relations, then, induce a relation between tags and concepts:

$$(2^{tagid}, conceptid)$$

to be precise, between sets of tags and concepts. What our experiments will attempt to determine is whether this induced relationship is a systematic, useful one that helps in analyzing data. Thus, the question is whether useful patterns appear in this relationship.

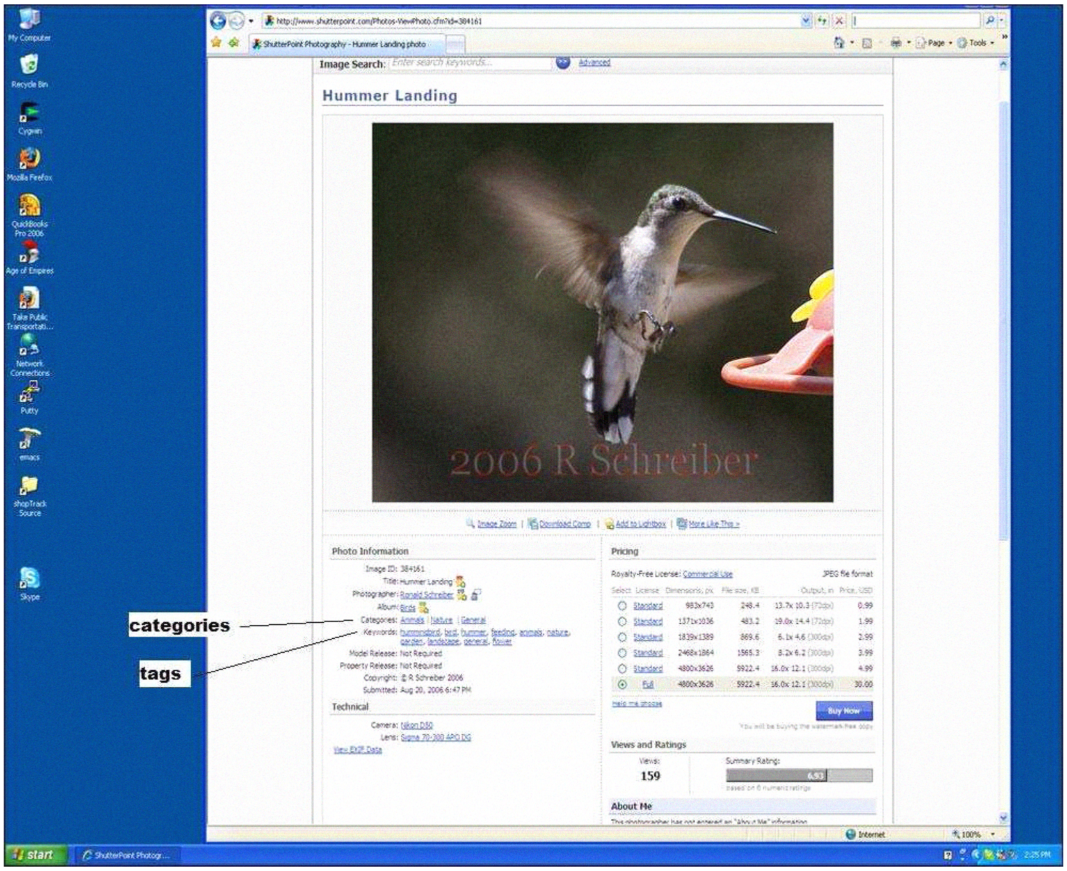
V. EXPERIMENTS

In this section, we describe two independent experiments that we carried out to determine the semantic value of tags. Both experiments have in common that they analyze data items (specifically, photographs from several photo web sites) by using *exclusively* user tags, and comparing this with existing taxonomies.

We carried out our experiments on photography web sites for several reasons. First, these sites contain the needed data, as they have large collection of photos and allow users to tag any and all of them. Second, the subject of the photos itself is very heterogeneous; therefore, there is very little risk that, when using the data for training, we will over-fit to any specific domain. Third, the sites are organized by the Web master in a hierarchy, giving a good ground truth as far as categorization is concerned. Finally, as image processing is extremely hard, relying on tags *only* is a natural alternative. Thus, these web sites are a perfect setup for our experiments.

A. Experiment 1: Concept Extraction

Our approach to concept extraction is to use supervised machine learning techniques to map a set of tags to concepts. We used tagged photographs from www.shutterpoint.com to evaluate our approach. A photo in Shutterpoint includes a set of tags and up to three categories (Figure V-A). The categories include concepts such as: 'still life', 'general', 'nature', and 'architecture'. The photo in Figure V-A has



Hummer Landing

Photo Information

Image ID: 384161
Title: Hummer Landing
Photographer: Ronald Schreiber
Album: Birds
Categories: Animals | Nature | General
Keywords: hummingbird, bird, hummer, feeding, animals, nature, garden, landscape, general, flower
Model Release: Not Required
Property Release: Not Required
Copyright: © R Schreiber 2006
Submitted: Aug 20, 2006 6:47 PM

Technical

Camera: Nikon D50
Lens: Sigma 70-300 APO DG
View EXIF Data

Pricing

Royalty-Free License: Commercial Use
JPEG file format

Select	License	Dimensions, px	File size, KB	Output, in	Price, USD
○	Standard	983x743	248.4	13.7x 10.3 (72dpi)	0.99
○	Standard	1371x1036	483.2	19.0x 14.4 (72dpi)	1.99
○	Standard	1839x1389	869.6	6.1x 4.6 (300dpi)	2.99
○	Standard	2468x1864	1565.3	8.2x 6.2 (300dpi)	3.99
○	Standard	4800x3626	5922.4	16.0x 12.1 (300dpi)	4.99
⊙	Full	4800x3626	5922.4	16.0x 12.1 (300dpi)	30.00

Help me choose

You will be buying the watermark-free copy

Views and Ratings

Views: 159

Summary Rating: 6.93
based on 6 numeric ratings

About Me

This photographer has not entered an "About Me" information

Fig. 1. Photograph in Shutterpoint. The photo of a humming bird is being tagged with keywords: hummingbird, bird, hummer, feeding, animals, garden, landscape, general, flower. It has also being associated with categories (concepts): Animals, Nature, and General

tags: hummingbird, bird, hummer, feeding, animals, garden, landscape, general, and flower. This photo has also being associated with concepts: Animals, Nature, and General. We used supervised classifiers to infer the concepts associated with photos based on their tags. To evaluate our approach, we compared inferred concepts with concepts assigned to the photos by their photographer.

Each photo is represented by its tags $\{t_1, \ldots, t_n\}$ and is associated with a single class. Each class represents the set of concepts that are associated with the photo. If photo p has tags $\{t_1, \ldots, t_n\}$ and concepts $\{c_1, c_2, c_3\}$, then $\{t_1, \ldots, t_n\}$ is associated with class $c_1 - c_2 - c_3$. Photo in Figure V-A is associated with class animals-nature-general. Our approach is to train a system to map a set of tags into a class representing a set of up to three concepts. The objective is to find a mapping that take tags $\{t_1, \ldots, t_n\}$ and returns class $c_1 - c_2 - c_3$.

We use weka's SMO classifier to train and predict concepts assigned to photos by their photographers. This classifier implements the sequential minimal optimization algorithm for training a support vector classifier [15]. Support vector classifiers represent input data as vectors in an n-dimensional space. This classifier finds maximum margin hyperplanes to separate input data into different classes. Maximum margin hyperplanes give the greatest separation between classes [15].

We train this classifier with pairs $< t, c >$ where t is a set of tags and c is a class representing up to three concepts. Then, we use the classifier to predict the class of a set of tags. The predicted class is a representation of a set of up to three concepts.

Results and Evaluation

To evaluate our approach, we used a set of photos from Shutterpoint. We partition this set into a training and a test set. Two thirds of these photos were assigned to the training set and the other third to the test set. We applied weka's SMO classifier on the training set and then on the test set to evaluate our approach.

The tags of photos in Shutterpoint were pre-processed to address some of the problems of uncontrolled vocabularies as follows. We retrieved a set of tags associated with photographs at `www.shutterpoint.com`. Then, the tags are processed in steps, as follows. We remove extraneous characters in tags such as punctuation symbols and noise words such as 'and', 'my', and 'you'. Then we fixed the spelling errors by replacing misspelled tags with the first suggestion that the Microsoft Word spelling checker gives. Then, we represent synonyms by a single tag. We use WordNet to replace all tags in a synset by the tag that was processed first. For example, Suppose that the

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330

first photo processed has tags: 'car', 'antique', and 'France' and the 10th photo has tags: 'auto', 'sports', and 'speed'. Because 'car' and 'auto' belong to the same synset and 'car' appeared first, then tag 'auto' in tenth photo is replaced by 'car'. Notice that any of the tags in the synset could have been selected to represent th synset. The objective is to have a single tag represent all the synonyms. Then, we reduce the tags to their root form by applying a stemmer.

Stage in pre-processing	Number of tags
Before pre-processing	37,416
After removing extraneous characters and noise words	36,710
After fixing spelling errors and reducing synonyms to a single tag	30,072
After stemming	11,456
After removing tags that appear only once in the entire set of photos	7,900

TABLE III
REDUCTION OF NUMBER OF DISTINCT TAGS DURING TAGS PRE-PROCESSING

Concept	Precision	Recall	F
Overall	*64%*	*64%*	*64%*
animals	81%	84%	82%
ocean	76%	81%	78%
botanical	75%	80%	77%
nature	70%	85%	77%
architecture	70%	73%	72%
religion-and-spirituality	77%	66%	71%
landscape	66%	76%	71%
night-shot	69%	65%	67%
macro	66%	67%	67%
travel	66%	61%	63%
transportation	67%	56%	61%
holiday	64%	49%	56%
urban-life	56%	50%	53%
rural-life	49%	43%	46%
abstract	53%	40%	46%
people-and-lifestyles	60%	35%	44%
backgrounds-and-textures	44%	37%	40%
fine-art	45%	32%	37%
general	36%	36%	36%
still-life	33%	27%	29%

TABLE IV
EVALUATION RESULTS OF EXTRACTING UP TO THREE CONCEPTS PER PHOTOGRAPH. OVERALL AND PER CATEGORY PRECISION, RECALL, AND F-MEASURE SORTED BY F-MEASURE

Our evaluation on a set of $24,998$ photos from Shutterpoint resulted in $16,665$ photos in the training set and $8,333$ photos in the test set. The number of distinct tags used in the entire set of photographs were $37,416$ and $7,900$ before and after pre-processing, respectively. As shown in Table III, the pre-processing reduced significantly the number of tags representing these photos.

We used weka's SMO classifier to train and predict concepts associated with photos in the test set by the photographer. The class predicted by the classifier is an encoding of the set of concepts associated with the photo. We decoded the class predicted by the classifier into a set of up to three concepts. Then, we compared the decoded set with the concepts assigned to the photo by its photographer.

We computed the overall precision, recall, and F-measure giving partial credit to partial correct predictions. To illustrate suppose that the classifier predicts class x-y-z for photo p whose photographer assigned concepts x and m. The number of true positives for p would be 1 (x), the number of false negatives would be 1 (m), and the number of false positives would be 2 (y and z). Then, the precision for p would be $\frac{1}{3}$ and the recall would be $\frac{1}{2}$. The overall precision, recall, and F-measure was computed by considering the aggregate number of true positives, true negatives, and false positives over all photographs, as follows: $Precision = (TP)/(TP + FP)$, $Recall = (TP)/(TP + FN)$, and $F = (2 \times precision \times recall)/(precision + recall)$. where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives.

To compute precision, recall, and F for each individual concept c, we set the number of true positives as the number of photographs assigned by their photographers to concept c such that the classifier predicted c as one of their concepts. Similarly, we set the number of false negatives as the number of photographs assigned by their photographers to concept c but the classifier did not predict c as one of their concepts. And, we set the number of false positives as the number of photographs whose predicted concepts included c but their photographer did not assigned c to them.

The overall precision (First row in second column of Table IV) reflects the percentage of concepts derived from the tags that were actually assigned to the photo by its photographer. The 64% precision indicates that approximately two out of three concepts derived from the tags were actually assigned to the photo by its photographer. The overall recall (First row in third column of Table IV) reflects the percentage of concepts assigned to the photo by the photographer that we were able to derive from the tags. The 64% recall indicates that approximately two out of three concepts assigned by the photographer to his/her photo were derived from the tags in the photo.

Second to fourth columns in Table IV show precision, recall, and F for individual concepts. We observe that our approach has good precision and recall when extracting from tags concrete concepts such as animals, ocean, and nature. On the other hand, the precision and recall decreases for vague concepts like general.

We also tried to predict one concept per photograph instead of up to three concepts. Each photograph associated with n concepts was represented in the training set as n instances, one for each concept. Each instance consists of the tags in the photo and a concept associated with the photograph. A photo with a set of tags t and concepts c_1, c_2, and c_3 would be represented in the training set as three pairs $< t, c_1 >$, $< t, c_2 >$, and $< t, c_3 >$. We used weka's SMO classifier to train the classifier and predict a single concept for each photo in the test set. We measured the overall precision and the precision for each category (Table V). The overall

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330

Concept	Precision
Overall	*80%*
animals	93%
botanical	90%
religion-and-spirituality	90%
holiday	88%
night-shot	86%
ocean	84%
macro	84%
architecture	82%
nature	79%
landscape	79%
travel	79%
transportation	79%
urban-life	77%
people-and-lifestyles	76%
rural-life	75%
abstract	74%
fine-art	64%
backgrounds-and-textures	63%
still-life	59%
general	45%

TABLE V
EVALUATION RESULTS OF EXTRACTING ONE CONCEPT PER PHOTOGRAPH. OVERALL PRECISION AND PRECISION FOR EACH CATEGORY SORTED BY PRECISION.

precision was 80% (Last columns in first row of Table V). This precision indicates that 80% of the photographs were correctly classified as one of the concepts assigned to the photo by the photographer. As for the case of predicting up to three concepts from the tags, the precision varies from concept to concept. Concrete concepts have better precision than vague concepts. Ninety three percent of photographs classified as *animals* based on their tags had *animals* as one of the concepts assigned to the photo by their photographer. On the other hand, 45% of photographs classified as *general* were correctly classified.

B. Experiment 2: Concept Similarity

In this experiment, we retrieved tags from items within a given concept, for a given Web site. That is, we chose a site s, a concept c within s's hierarchy, and extracted a random subset of tags attached to data items classified under c at s. The process is repeated at a different site s'. The concepts themselves were chosen at random but subjected to the criteria that some of them had to be common to two or more sites, while others had to be different. We then compared the resulting tag sets across concepts and across data sites. The hypothesis here is that similar concepts will exhibit a strong amount of similarity across sites, while different concepts will show low to none amount of similarity across sites (or even within the same site). To validate our hypothesis, we chose several sites that share some basic characteristics: they all offer photographs for sale, all of them classify photographs into basic concepts, and all of them also allow users to tag photographs. Note that these are basic requirements if the data in our experiment is to follow the framework outlined in section IV. The chosen sites were Shutterpoint (`www.shutterpoint.com`) and FeaturePics (`www.featurepics.com`).

Formally, the approach can be characterized as follows. After forming the $(conceptid,\ tagid)$ relation for each site, compare across sites and choose concepts that appear in at least two sites, as well as different concepts. Then the resulting relations are grouped by $conceptid$ to get a bag (multiset) of tags for each concept. Finally, we compare the bags of tags for each pair of concepts. Since the number of tags for a given concept can vary from concept to concept and from site to site, and the amount of (non) overlap may depend on the number of tags, we simply determine a similarity measure and do a series of pairwise comparisons. We then compare the numbers obtained for similar concepts and the numbers obtained for dissimilar concepts. A good measure is one that give higher numbers to similarity *and* low numbers to dissimilarity, so that it can act as a *discriminator*.

We use as our similarity measure the well-known Jaccard function: given two sets A and B, this is simple $\frac{|A \cap B|}{|A \cup B|}$. However, the function needs to be adapted to our environment, since we are dealing with bags (multisets), and we consider that a certain amount of noise (represented by tags which occur infrequently) is present. One way to adapt the measure is to use multiset intersection and union. We used this approach (slightly altered, see formula below) in a measure that is called Jaccard1 in what follows. However, such adaptation, while respecting the semantics of multisets, does not deal with noise, something which is especially important in unrestricted settings. Thus, we also devise several other measures.

Let S, T be bags (multisets) of tags, 'a' a tag, and fr(a,S) the frequency (number of occurrences) of 'a' in S. A *tag multiset* can be represented as a set of pairs $(a,\ fr(a,S))$. Clearly $|S| = \Sigma_{a \in S} fr(a,S)$. We define $Pr(a,S)$, the normalized frequency of 'a' in S, as $\frac{fr(a,S)}{|S|}$. A *normalized* tag set is one where the frequency of each element has been replaced by the normalized frequency, that is, a set of pairs $(a,\ Pr(a,S))$. A *truncated* tag set is one where certain elements have been eliminated by giving a cut-point. In a non-normalized tag set, the cut-point gives a minimum raw frequency the element must have; in a normalized tag set, the cut-point gives a minimum relative frequency. Given cut-point α, $Tr(S,\alpha)$ denotes the truncated set; hence $Tr(S,\alpha) = \{a \in S \mid fr(a,S) > \alpha\}$ for non-normalized bags; and $Tr(S,\alpha) = \{a \in S \mid Pr(a,S) \geq \alpha\}$ for normalized bags. The intuition behind truncated sets is that, since most collection of tags exhibit a power-law like distribution, we can eliminate the long tail of the distribution and concentrate on the tags that appear often, which usually are a small number.

The following measures were used in our experiment:

Jaccard1, as explained above, is simply the standard Jaccard with bag semantics for intersection and union, on the raw (non-normalized, non-truncated) tag set.

$$Jaccard1 = \frac{\Sigma_{a \in S \cap T} min(fr(a,S), fr(a,T))}{\Sigma_{a \in S \cup T} fr(a,S) + fr(a,T)}$$

Note that we take the minimum for the intersection, as usual, but the sum (as opposed to the maximum) for union. As a result, our measure is a bit stricter (results in a smaller number) than the standard measure.

- *Jaccard2* is again standard Jaccard with bag semantics over a truncated, non-normalized set.

$$Jaccard2_{\alpha} = \frac{\Sigma_{a \in Tr(S,\alpha) \cap Tr(T,\alpha)} min(fr(a,S), fr(a,T))}{\Sigma_{a \in Tr(S,\alpha) \cup Tr(T,\alpha)} fr(a,S) + fr(a,T)}$$

The parameter α refers to the cut-point used for the truncation. Note that $fr(a,S) = fr(a,Tr(S,\alpha))$ whenever $a \in Tr(S,\alpha)$. This is equivalent to declaring $fr(a,S) = 0$ whenever $fr(a,S) \leq \alpha$.

- *Jaccard3* is the standard Jaccard over a normalized tag set, using the normalized frequency of elements as a *weight* in calculating the intersection (union).

$$Jaccard3 = \frac{\Sigma_{a \in S \cap T} min(Pr(a,S)Pr(a,T))}{\Sigma_{a \in S \cup T} Pr(a,S) + Pr(a,T)}$$

- *Jaccard4* is standard Jaccard on a truncated, normalized set.

$$Jaccard4_{\alpha} = \frac{\Sigma_{a \in Tr(S,\alpha) \cap Tr(T,\alpha)} min(Pr(a,S), Pr(a,T))}{\Sigma_{a \in Tr(S,\alpha) \cup Tr(T,\alpha)} Pr(a,S) + Pr(a,T)}$$

The big difference between truncating a normalized and a non-normalized tag set is that for normalized sets, one cut point can be given that can be used for all sets, while for non-normalized set it may be necessary to adjust the cut point from set to set.

Results and Evaluation

The first set of results was obtained by comparing categories Children in `Shutterpoint.com` and `FeaturePics.com`, as well as category Military from `Shutterpoint.com` to Children from `FeaturePics.com`. We used all four measures to determine if any one of them was significantly better than the others. The size and amount of duplication in the tag set are shown in table VI. The last row includes the number of tags in the truncated, non-normalized set; the number in parenthesis is the cut-point (α) used. Note that the cut-point is rather small; only tags appearing a single time (twice for the Children category) are thrown away.

Site	**Shutterpoint**	**FeaturePics**	**Shutterpoint**
Concept	**Children**	**Children**	**Military**
Nr. of tags	1956	18352	4750
Nr. of unique tags	1658	2011	1773
Nr. of top tags	251 (1)	852 (2)	852 (1)

TABLE VI
TAG CHARACTERISTICS

The results are shown in the table of Figure VII. For jaccard4, a universal cut-point of 0.001 was experimentally established as a good threshold to get rid of the distribution's tail. Focusing on the ability of the measures to distinguish between similar and different concepts, Jaccard3 and Jaccard4 are the best measures. This can be attributed to the fact that truncating a normalized tag set does a good job of getting rid of the tail of the distribution, which includes all tags that can be considered as noisy. To make clear the discriminating power, additional results are presented in the table of Figure VIII. There, several pairs of categories, sometimes similar and sometimes dissimilar, but each from a different web site, are compared. Thus, for instance, the first (top) row presents the result of comparing similar categories "Children" from two sites, while the last (bottom) row compares dissimilar categories "Children" and "Military", each from a different site. We expect the numbers for similar categories to be consistently higher then the numbers corresponding to different categories. The results bear this: the *lowest* similar result shows a 2.15 ratio to the *largest* dissimilar result (measure is more than twice as strong when categories are semantically similar). Also, absolute numbers are consistent with previous experiment: semantically similar categories give a measure between 0.25 and 0.30; semantically unrelated, about .1.

Measure	Children-Children	Children-Military
base	.0389	.017
Jaccard1	.0457 (709 tags)	.038 (399 tags)
Jaccard2	.020 (157 tags)	.030 (147 tags)
Jaccard3	.278 (709 tags)	.119 (399 tags)
Jaccard4	.259 (157 tags)	.105 (147 tags)

TABLE VII
SIMILARITY MEASURES

Categories Compared	Score
Children-Children	0.259
Transportation-Transportation	0.246
Military-Transportation	0.114
Children-Transportation	0.062
Children-Military	0.105

TABLE VIII
MULTIPLE PAIRWISE COMPARISONS

VI. CONCLUSION AND FURTHER RESEARCH

In this paper, we attempt to determine the amount of semantics that tags carry about data they are used with. To achieve this, we design and implement two experiments where the tasks are widely seen as involving some aspect of data semantics. In particular, we study the usability of tags for concept extraction and determining equivalence relations between concepts based on the tag sets associated with these concepts.

Our first experimental results on tagged photographs from the stock photography Web site `www.shutterpoint.com` showed that our approach can extract up to three concepts with an overall precision of 64% and recall of 64%. The precision and recall of concept extraction depends on the concepts associated with the photographs. Extraction of concrete concepts have higher precision and recall than vague concepts. Our approach was able to extract concrete concepts such as *animals* with precision of 81% and recall of 84%, while for

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330

vague concepts such as *general*, the precision and recall was just 36%. When reducing the number of concepts to extract from up to three to one, our approach was able to extract a single concept from the photograph's tags with an overall precision of 80%. The precision of extracting a single concept from the photo's tags varied from concept to concept. This precision ranged from 45% to 93%.

Our second experimental results used tags from two sites, `www.shutterpoint.com` and `www.featurepics.com`. We compared tag sets from different sites, and we showed that tag sets constitute good discriminants of *semantic similarity*. Experimental results showed that our similarity metrics on similar concepts were significantly higher than on dissimilar concepts. Taking into account that each Web site has a different audience (and we think it's highly likely that the audiences have little overlap), and that the concepts come from different taxonomies with different characteristics (one is flat, the other one has several levels), we consider these results highly significative.

All in all, our experimental results with real-life Web data show that tags are indeed useful for these tasks. Our results empirically validate some past work by other researchers, as our experiments involve more Web sites than such work, usually limited to one web site. Based on this evidence, we believe it is fair to say that tag sets carry a significant amount of semantics.

However, it is clear that much work remains to be done. In further research, we plan to generalize these results by comparing more categories across more Web sites, and to carry out a more fine-grained analysis to try to establish how much data (how many tags) are needed to establish stable results. Also, we plan to expand the experiments across domains, that is, compare tags from completely different sites (say a news site and a photography site) on which common categories can be found, to see if tags still carry semantics in such a situation. Note, though, that in order to have some common categories, the domains of the sites cannot be completely disjoint. Also, note that as far as *subject matter* is concerned, our chosen web sites were quite heterogeneous; therefore, we believe that our results also have a degree of robustness -but obviously would like to expand the data sets considerably before making any definitive claims. We also plan to apply this research in a novel architecture for information integration that SET is currently developing.

ACKNOWLEDGMENTS

We thank Dr. Todd Hughes for his feedback and guidance in the development of the ideas presented in this paper. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Army Aviation and Missile Command.

REFERENCES

[1] H. S. Al-Khalifa and H. C. Davis, "Measuring the semantic value of folksonomies," in *Proc. of International Conference on Innovations in Information Technology*, Nov 2006.

[2] L. Zhang, X. Wu, and Y. Yu, "Emergent semantics from folksonomies: A quantitative study," *Journal on Data Semantics VI*, 2006.

[3] X. Wu, L. Zhang, and Y. Yu, "Exploring social annotations for the semantic web," in *WWW '06: Proceedings of the 15th international conference on World Wide Web*. New York, NY, USA: ACM Press, 2006, pp. 417–426.

[4] A. Mikroyannidis, "Toward a social semantic web," *Computer*, vol. 40, no. 11, pp. 113–115, 2007.

[5] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, April 2006.

[6] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?" in *Proc. of ACM International Conference on Web Search and Data Mining*. ACM, Feb 2008.

[7] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. ACM Press, 2007, pp. 501–510.

[8] S. Xu, S. Bao, Y. Cao, and Y. Yu, "Using social annotations to improve language model for information retrieval," in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 1003–1006.

[9] P. Mika, "Ontologies are us: A unified model of social networks and semantics," in *International Semantic Web Conference*, ser. Lecture Notes in Computer Science, International Semantic Web Conference 2005. Springer, November 2005, pp. 522–536.

[10] A. Plangprasopchok and K. Lerman, "Exploiting social annotation for automatic resource discovery," in *Proc. of AAAI workshop on Information Integration from the Web*. AAAI, Apr 2007.

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

[12] M. Kato, H. Ohshima, S. Oyama, and K. Tanaka, "Can social tagging improve web image search?" in *Proc. of Conference on Web Information Systems Engineering*, Sep 2008.

[13] A. Y. Halevy, N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka, "Enterprise information integration: successes, challenges and controversies," in *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2005, pp. 778–787.

[14] B. A. Huberman, "The structure of collaborative tagging systems," in *HP Labs*, 2007.

[15] I. Witten and E. Frank, *Data Mining - Practical Machine Learning Tools and Techniques*. Elsevier, 2005.

Digital Object Identifier: 10.4108/ICST.COLLABORATECOM2009.8330
http://dx.doi.org/10.4108/ICST.COLLABORATECOM2009.8330