Harvesting Geospatial Knowledge from Social Metadata

Suradej Intagorn
USC/Information Sciences Institute
intagorn@isi.edu

Anon Plangprasopchok
USC/Information Sciences Institute
plangpra@isi.edu

Kristina Lerman
USC/Information Sciences Institute
lerman@isi.edu

ABSTRACT

Up-to-date geospatial information can help crisis management community to coordinate its response. In addition to data that is created and curated by experts, there is an abundance of user-generated, user-curated data on Social Web sites such as Flickr, Delicious, and Google Earth, that can be used to harvest knowledge to solve real-world problems. User-generated, or social, metadata can be used to learn concepts and relations between them that can improve information discovery, and data integration and management. We describe a method that aggregates social metadata created by thousands of users of the social photo-sharing site Flickr to learn geospatial concepts and relations. Our method leverages geotagged data to represent and reason about places. We evaluate learned geospatial relations by comparing them to a reference ontology provided by GeoNames.org. We show that our approach achieves good performance and also learns useful information that does not appear in the reference ontology.

Keywords

Geospatial, metadata, ontology, geo-ontology, Social Annotation, Social Web, geotagging, Flickr.

INTRODUCTION

When disaster strikes, the humanitarian relief community needs quick access to geospatial data to assess damage and coordinate relief efforts. Much of the relevant data is available online, in the form of geo-referenced images, geoRSS feeds, and kml files containing map overlays and place-marks. Recent progress in information integration has led to novel applications that allow decision makers to accurately combine heterogeneous geospatial information sources, e.g., satellite imagery with maps (Michalowski et al., 2007). However, discovering relevant geospatial information, assessing its quality, and annotating it to make it available for integration is still a laborious manual process (Kavouras et al., 2006). The solution advocated by the data management community is to semantically annotate data with terms from a predefined ontology. This means that a group of experts has to agree on common semantics, expressed through an ontology, and then have data providers annotate content with concepts from this ontology. While this approach works well for homogeneous, centralized organizations, it does not scale to the dynamic heterogeneous environment (Euzenat and Shvaiko, 2007), such as the Web.

Information and knowledge production is no longer solely within the purview of experts and professionals. In many areas of intellectual inquiry, vast armies of lay volunteers are creating, publishing, and annotating rich content on Social Web sites such as *Flickr*, *Delicious*, *OpenStreetMap*, among many others. The information people create while organizing and using content and interacting with other people is called social metadata. Social metadata includes *tags*, or labels people use to describe content, *relations* that are used to hierarchically organize content or metadata, as well as geographic coordinates attached to content, known as *geotags*. Although social metadata is freely generated and uncontrolled, it reflects how a community organizes knowledge, including geospatial knowledge. Integrating such ‘folk knowledge’ with that created by experts will greatly enhance geospatial information integration. Take as example, the recent Station fire in La Cañada, a small city on the outskirts of Los Angeles. People uploaded many images to the social photosharing site Flickr detailing the progress of the fire, its destruction, and the pollution it created. While many of these images were “correctly” tagged with ‘La Cañada’ and ‘wildfire’, others were “incorrectly” tagged with ‘Los Angeles’ and ‘wildfire’. By “incorrect” we mean that no expert would have created those tags, although many lay people did.

There are several advantages to leveraging social metadata to learn geospatial concepts and relations. First, social metadata is *distributed* and *dynamic* in nature (Golder and Huberman, 2006); therefore, more likely to stay *complete* and *current* than formal ontologies created by groups of experts. More importantly, it is closer to the ”common knowledge” shared by a community . `GeoNames.org` (http://geonames.org), for example, is a geographical database containing millions of geographical names, formally categorized within a taxonomy. This database is maintained by a small community of experts from several different countries. Unlike many Flickr users who place La Cañada in the “greater Los Angeles”, `GeoNames` does not. Searching Flickr using the query terms suggested by `Geonames` only would not retrieve all relevant information about the Station fire.

Community-generated knowledge that is automatically extracted from social metadata could, therefore, complement and provide different perspective to existing geospatial ontologies created by experts (Keating and Montoya, 2005). Another advantage of this approach is that learned geospatial concepts are directly linked to content, enabling diverse geospatial data to be more easily used within applications, integrated and aligned across domains. Geospatial applications that rely on formal ontologies would first need to map user-created content to the ontology before such content could be used within the application. In this paper we describe a method to extract geospatial knowledge about places and relations between them from social metadata created by large numbers of users on the Social Web. First, we describe the types of available social metadata and illustrate with examples from the social photo-sharing site `Flickr`. Next, we describe our approach to identifying and representing places, and learning relations between them. We apply the proposed approach to metadata extracted from `Flickr` and compare its performance to the current state-of-the-art relation learning method. We conclude with review of relevant research and future research directions.

SOCIAL METADATA

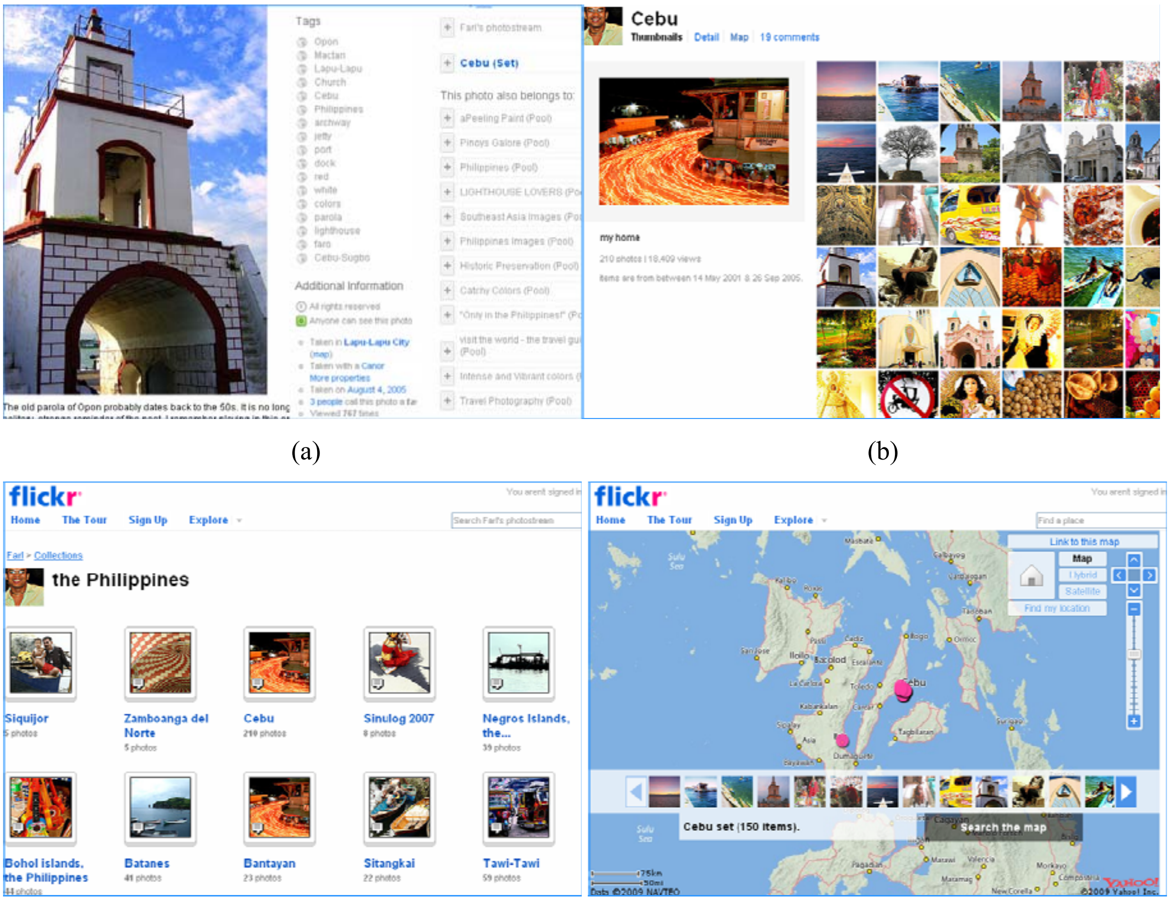
Tagging has become a popular method for annotating content on the Social Web. When a user tags an object, be it a Web page on `Delicious`, a scientific paper on `CiteULike`, or an image on `Flickr`, the user is free to select any keyword, or tag, from an uncontrolled personal vocabulary to describe the object. In addition to tags, some social Web sites, such as `Delicious`, and `Flickr`, have recently begun to provide a feature enabling users to organize content hierarchically. While the sites themselves do not impose any constraints on the vocabulary or the semantics of the hierarchical relations used, in practice users employ them to represent both subclass relationships (‘paris’ is a kind of ‘city’) and part-of relationships (‘yosemite’ is a part of ‘california’) (Plangprasopchok and Lerman, 2009). We claim that these diverse forms of social metadata offer a rich source of evidence for learning how people organize knowledge. Although we describe the types of metadata and illustrate with examples from `Flickr`, similar functionality is offered by other Social Web sites.

Tags: Figure 1 (a) shows an image on `Flickr`, along with metadata associated with it. Tags that describe this image include such useful descriptors as image’s subject (“lighthouse” or “parola” in Tagalog), features (“archway”, “jetty”, “dock”) and colors (“white”, “red”), in addition to where the image was taken (“Lapu-Lapu”, “Cebu”, “Philippines”). The image was also added to public groups (followed by the word “(Pool)”) devoted to lighthouses, historic preservation, travel, Philippines, and several others.

Sets and collections: `Flickr` allows users to group photos in folder-like *sets*, and group sets in *collections.*[1] Both sets and collections are named by the owner of the image, and a photo can be part of multiple sets. `Flickr` does not enforce any rules about organizing photos in sets and collections or naming them. While some users create multi-level hierarchies, the majority create shallow hierarchies consisting of collections and their constituent sets. The image in Figure 1 (a) was grouped with other images taken around the Philippine province of Cebu in an eponymous set (Figure 1 (b)). This and sets devoted to other places around Philippines were grouped together in a collection called “the Philippines” (Figure 1 (c)). Other collections created by this user on `Flickr` include “Indochina”, “East Asia”, “Tanzania”, “US”, and “Norway”.

[1] The collection feature is limited to paid “pro” users. Pro users can also create unlimited number of photo sets, while free membership limits a user to three sets.

Geotags: In addition to keywords, users can attach geospatial metadata to photos in the form of geographic coordinates. As an interesting observation, geotagging emerged spontaneously on `Flickr`, as users began to tag their photos with “geo:lat” and “geo:long.” `Flickr` later introduced an integrated geotagging/mapping feature, which enables users to display their images on a map or georeference them with a single click. Figure 1 (d) shows images (purple dots) in the “Cebu” set displayed on a map.



(a) (b)

(c) (d)

Figure 1 Examples of data and metadata created by a Flickr user. (a) Tags assigned to an individual image (geotags are not shown), (b) images in the set “Cebu”, (c) sets in a collection called “the Philippines” created by the user, (d) geotagged images in the “Cebu” set displayed on the map.

LEARNING ABOUT PLACES FROM SOCIAL METADATA

We believe that social metadata, specifically metadata linked to geographic coordinates, provides a valuable source of evidence for learning about places. Acquiring accurate geospatial knowledge, however, presents several challenges. First, people describe their content using highly idiosyncratic vocabulary. Second, the keywords selected as tags or set names may be prone to ambiguity (same keyword means different concepts) and synonymy (different keywords refer to the same concept) (Mathes, 2004; Golder and Huberman, 2006). Third, different people may have different levels of expertise and expressiveness (Golder and Huberman, 2006). A knowledgeable user, for example, is more likely to use more specific terms to express finer-grained concepts, while a less knowledgeable user will use more general terms. An expressive user will annotate content with many terms, but most users will use very few, on the order of 4-7, terms for annotation. In addition, people indiscriminately combine terms referring to different facets of data (Mathes, 2004; Rashmi, 2005), e.g., “USA 2006” for photos of travel in the United States in 2006. Finally, data may not be homogeneously distributed, with more geo-tagged data in heavily populated places and popular tourist destinations.

We describe an approach for aggregating geo-tagged metadata on `Flickr` to learn about geographic concepts, or places, and relations between them, deals with some of the challenges described above. Our approach, shown in Figure 2, consists of two main steps: (i) Recognizing places, e.g., ‘cebu’ and ‘philippines’, and (ii) using geospatial subsumption to learn part-of relations between places, e.g., ‘cebu’ is part-of ‘philippines’. Each step consists of important sub-steps, which are described in greater detail below.

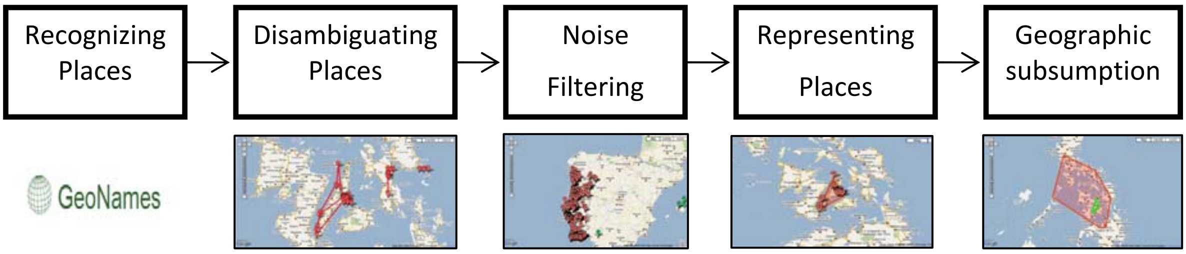


Figure 2 Flowchart of the learning approach

Recognizing Places

The first step is to identify places from social metadata and obtain a representative set of points. Specifically, we define a place as an association between the name of that place and a collection of geographic points. We assume that textual metadata, such as tags or sets, represent specific concepts, including geospatial concepts, and derive *place names* from them. The points are obtained from the geographic coordinates (latitude, longitude pairs) of geotagged photos associated with these names.

Since names can refer to any concept, not necessarily a place, we have to filter out “non-place” names. Recently, Rattenbury and Naaman (2009) proposed an approach to automatically detect tags associated with places on Flickr by analyzing the spatial distribution of the coordinates of all photos associated with that tag. If the distribution is highly localized, the tag is determined to be a place tag. However, this approach is likely to suffer from the challenges of noise, ambiguity and idiosyncratic naming conventions. Instead, we use set names as an alternative evidence for a place. As shown (Plangprasopchok and Lerman, 2009), using sets to represent concepts rather than tags is more robust with respect to the challenges of learning from social metadata.

We use *GeoNames* as a reference set to help us recognize place names. Although such a reference set may not contain a comprehensive list of places, in this paper we simply check whether a particular name in *GeoNames* is contained within a set name. We normalize data by lowercasing both Geonames and set names.

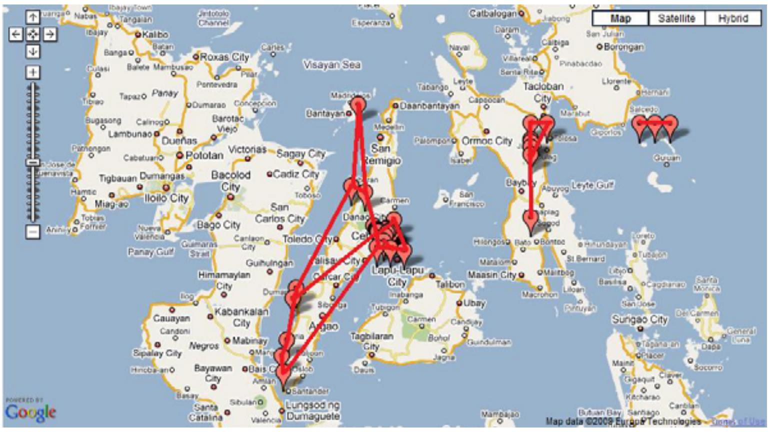


Figure 3 Creating a graph of points associated with place ‘cebu’

Disambiguating Places: A given place name can be ambiguous, and therefore, consist of a mixture of points from different places that share the same name. For example, ‘victoria’ can refer to a place in Canada or Australia, among others. Similarly, ‘cambridge’ can be found in United Kingdom and United States. The natural solution to this challenge is to cluster points with the same name. Specifically, we assume that points associated with the same place are closer to each other than those associated with other places. A computational method that clusters geographic points by their location can implicitly resolve place name ambiguity by separating points of one place from the others. To cluster points for a given name, *n*, we first obtain all points from all sets with name *n*. Each point, *v*, has latitude and longitude, which will be used to compute Euclidian distance between the points. However, there are places that are composed of non-contiguous subregions, which are distant from each other, e.g., Hawaii and Alaska are part of United States, even though they are far from the mainland of United States. To link distant regions to a given place, we exploit constraints imposed by photo sets on `Flickr`. Specifically, *we assume that points from the same set belong to the same place*. If there are points in two clusters that belong to the same set, these two clusters would then belong to the same place. We capture these constraints between points in a graph and then analyze the graph to discover distinct places.

We create $G1_n = (V_n, E_n)$, an undirected graph of points associated with a geo-name *n*. Vertices V_n are points corresponding to geo-tagged photos in sets with name *n*, and E_n are the edges between vertices. Let s_{vi} be the set index of vertex $v_i \in V_n$. An edge between two photos is created if and only if $dist(v_i,v_j) < \tau$ or $s_{vi} = s_{vj}$. Here, $dist(v_i,v_j)$ is the Euclidean distance between points v_i and v_j and τ=500 km. Figure 3 shows the graph for points associated with the place name 'cebu'. After we create the graph in this manner, we find its maximally connected components (Hopcroft and Tarjan, 1973), with each component corresponding to a different sense of the name. Thus, points associated with the name 'victoria', for example, will be divided into two sub-graphs, one of which is located in Canada and the other one in Australia.

After disambiguating places, we cluster the points in each disambiguated place again using the distance condition only. This helps identify disjoint regions associated with a place, for example, 'usa' can be composed of three distinct clusters corresponding to continental US, Hawaii, and Alaska. Sometimes a single contiguous place is represented by more than one cluster. In our data set, 'canada' is represented by two clusters, one for Eastern Canada and one for Western Canada. We find that these clusters correspond to heavily populated or traveled places. Our approach, therefore, has to represent places as non-contiguous regions.

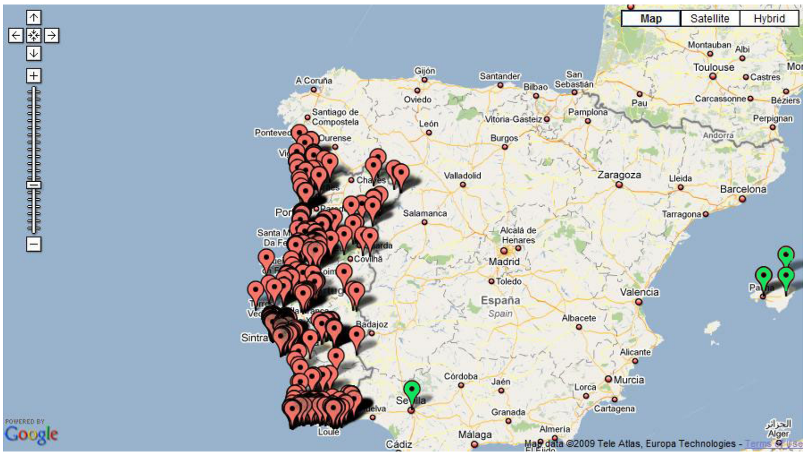


Figure 4 Example of noise filtering. Points are associated with sets containing name 'portugal'. Points in countries other than Portugal (in green) are filtered out.

Noise Filtering: As mentioned earlier, social metadata created by diverse users is noisy. Noisy data can significantly distort our representation of places and degrade the performance of the learning algorithm. For example, there are photos taken at Los Angeles International Airport (LAX) that appear in a set "Australia." Any representation of 'australia' that includes parts of Los Angeles will lead to inaccurate relations between Australia and other places. In this section, we describe an approach to filter out noisy data.

As illustrated in the example above, noise can appear due to idiosyncratic tagging by individual users. This leads us to identify two characteristics of noise: (a) it is very different from other similar data (LAX points are very far from the other 'australia' points), and (b) it is created by a small number of users (it is highly unlikely that more than one user added points around LAX to a set named "Australia"). Let U_{ci} be number of users who geo-tagged photos in cluster *i*. We filter the noisy cluster out if $U_{ci} < \kappa$. In our experiments, we set κ=2. That is, if a given cluster contains points from only one user, it is very likely to be noise.

Noise can also lead to errors in estimating boundary of a place. For example, some of the points in sets called 'canada' are actually located in the United States, because people often include US border regions in their travel to Canada. The result is that the place 'canada' will include points in the United States. Most of them will occur as a single point or small group of points. We detect this type of noise by its locality. In our implementation, we average distance of a point to its K-nearest points and filter out *N*% of farthest points. In our experiments, N=5.

Figure 4 shows the points associated with a place 'portugal'. Most of the points are located in Portugal, although a few others, shown in green, are in other countries, such as Spain and Italy. The method described in this paper filters out these points.

Reasoning about Places

After filtering out noise and non-place concepts, points from the remaining concepts exhibit high locality. Moreover, places are expressed at different levels of granularity, from continents and countries to cities and parks. In this section we discuss our scheme for representing and reasoning about places. This scheme allows us to use geospatial subsumption to learn relations between places, e.g., that 'california' is part of 'united states' and 'cebu' is part of 'philippines', but also formally "incorrect" relations that reflect folk knowledge, e.g., 'la cañada' is part of 'los angeles'.

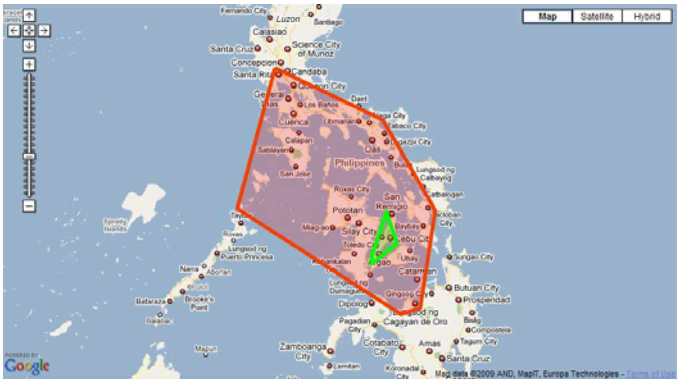


Figure 5 Convex hulls created from points representing 'cebu' (green) and 'philippines' (red). Note that 'cebu' is subsumed by the 'philippines'.

Representing Places: In current work, we represent each place as one or more convex hulls (multiple convex hulls). Although, there are drawbacks to this approach, e.g., convex hulls cannot guarantee correctness of the boundary, we use this representation for the initial study reported in the paper both to validate the concept of our approach and to provide a baseline for future work. In the future, we plan to use arbitrary polygons to represent places in order to improve the accuracy of learning. To identify convex hulls, we use the approach proposed by Hopcroft (1973)[2]. For each cluster obtained by our method, we estimate its boundary as a convex hull of its points. One reason to use multiple convex hulls instead of one is that some places, such as US, may be composed of non-contiguous regions, and representing such a place as a single convex hull will include regions between unconnected regions, e.g., Canada. The other reason to use multiple convex hulls is from the sparseness of data. For example, Canada has many points on the East and West coasts, but relatively few points in the middle of the country. Although the region is separated into two convex hulls, which may lead our method to learn incorrect relation for some places in the middle of Canada, we sacrifice these instead of learning incorrect relations caused by a poor estimate on the region.

Figure 5 shows the convex hulls created from the points representing 'cebu' (green) and 'philippines' (red). Note that the red region contains, or *subsumes*, the green region, from which we may reasonably conclude that 'cebu' is a part of 'philippines', or that 'philippines' is a parent of 'cebu'. We use geographic subsumption to discover these relations in the data.

Geographic Subsumption: We adapt probabilistic subsumption method (Sanderson and Croft, 1999; Schmitz, 2006) to determine whether one place subsumes another. We use the boundary of the convex hull and its points to determine geographic subsumption relations. Basically, we determine the fraction of an area of one place that is contained within the boundary of another place. If most of the area of the former is within the boundaries of the latter, we say the latter subsumes the former. More precisely, we say that place A subsumes place B if "most" of B is contained within the boundary of A, but not vice versa. As in Schmitz (2006), A subsumes B if $p(\mathrm{A}|B) >= t$ and $p(B|A) < t$, where t is a predefined threshold. $p(B|A)$ is estimated from $Area(A \cap B)/Area(\mathrm{A})$, and $p(A|B)$ is estimated in similar manner, where $Area(A)$ is a function that returns the area of A, and $Area(A \cap B)$ returns the area of intersection of A and B.

EVALUATION

We used the `Flickr` API to retrieve the names of members of seventeen public groups devoted to wildlife and nature photography. We then used a Web page scraping tool to retrieve sets created by these users. We retrieved

[2] We use its implementation in JTS Topology Suite: http://www.vividsolutions.com/jts/jtshome.htm.

a total of 166,526 sets from 7,618 pro users, and also the tags and geotags from images in these sets, which yields 1.3 millions of geographical points (photos) in total.

We collected all points associated with each place name by identifying photos that are contained in sets whose name matches a geoname in *GeoNames.org*. We used substring matching to match the geoname to set name. The geotags of these photos then become our points. We identified 1,774 geographic concepts in the data set and used associated points to create regions. Of these concepts, there are 610 concepts about the North America continent.

We compare our approach to tag-based probabilistic subsumption described in Schmitz (2006). This baseline method computes *p*(*B*|*A*) from the co-occurrence of tags *A* and *B*: i.e., *p*(*B*|*A*)=*Frequency*(*A*,*B*)/*Frequency*(*A*), where *Frequency*(*A*) is the number of photos tagged with *A*, and *Frequency*(*A*,*B*) is the number of photos tagged with *A* and *B*. To collect data for this baseline, we queried *Flickr* to find the number of images that were tagged with keyword *A* and two keywords *A* and *B*. The keywords were geonames that matched set names in our data set. Unfortunately, since the baseline approach requires us to invoke *Flickr*'s webservice to obtain a co-occurrence count of each tag pair, it is infeasible for us to obtain all counts of the entire data set (which requires 1,774 choose 2 requests). Instead, we collect tag coocurrence statistics for photos on the North American continent (610 choose 2 requests).

As shown in Figure 6 below, the number of learned relations has the same trend as to the F-score in our approach. This is because many parent places geospatially subsume child places completely. Consequently, the change of threshold value will not affect to the number. On the other hand, the number of relations in the baseline approach decreases to zero as the threshold increases because in many cases, users do not specify parent and child tags in the same photo, which lessens a chance that the parent subsumes the child tag. For example, the tag 'miami' has been used 645,512 times and 'florida' 2,763,640 times. Hence, the co-occurrence frequency between them is only 187,561 times. At the threshold > 0.3, the relation 'miami' is part of 'florida' cannot thus be learned.

We evaluate the subsumption relations learned by our method and the baseline automatically. Basically, our automatic evaluation compares learned relations to the existing hierarchy in the reference set *Geonames.org*. If the relation exists in *Geonames*, the automatic evaluator marks the result as correct; otherwise, it is incorrect. In this evaluation, we compute precision as the number of relations marked correct automatically (AC) divided by the number of learned relations. Recall is computed as AC divided by the number of *Geonames* relations, whose place names appear in the Flickr dataset (matched to some set names).

Here, we report how the F-measure of the learned relations changes with the subsumption threshold *t*. F-measure is the harmonic mean of precision and recall. All experiments vary the threshold from 0 to 1 in steps of 0.01. As shown in Figure 7, F-score of our approach reaches a maximum at around 0.627 and is relatively constant over a wide range of thresholds. For the North America continent, our approach reaches maximum F-score at 0.5047 and produce similar trends; while the baseline reaches maximum F-score at 0.3551.

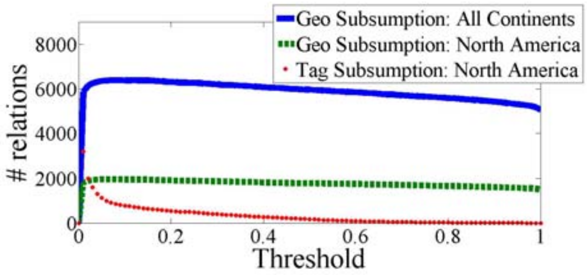


Figure 6. Numbers of geospatial relations induced by the proposed approach and the baseline at different values of the subsumption threshold, *t*.

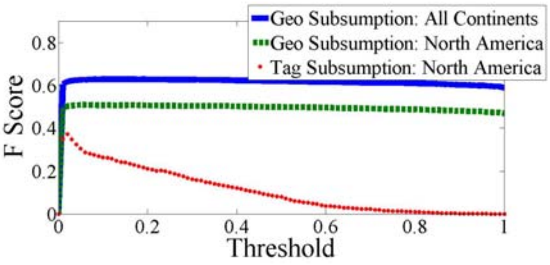


Figure 7: An automatic comparison against the reference set on F-score between the proposed approach and the baseline at different values of the subsumption threshold, *t*.

Our approach appears to be insensitive to the threshold *t* because many parent places are bigger than child places and often completely contain the child place. For examples, *p*('alcatraz' | 'usa') = 0 and *p*('usa' | 'alcatraz') = 1. For any *t* > 0.0, 'usa' will geographically subsume 'alcatraz'.

Child	Parent
anaheim	la
ballard	puget sound
brandywine park	wilmington
bronx	new york city
bronx zoo	new york city
bruce peninsula	georgian bay
burbank	la
burbank	los angeles
cabo san lucas	los cabos
cabrillo	san diego
chinatown	los angeles
coney island	new york city
crescent beach	nova scotia
dayton	new york
discovery park	seattle
disneyland resort	disneyland
eastern market	detroit
elkhorn slough	monterey
eureka	victoria
georgia aquarium	atl
pasadena	los angeles
disneyland	la
golden gate bridge	san francisco bay
greenfield	new york
griffith park	la
griffith park	los angeles
hollywood	la
la jolla	san diego
malibu	los angeles
pasadena	la
pearl harbor	oahu
queen anne	seattle
san diego wild animal park	san diego
san diego zoo	san diego
santa monica	la
santa monica	los angeles
sea world	orlando
times square	new york city
union square	manhattan
university of south florida	tampa
university of washington	puget sound
webster park	rochester
lake eola	orlando
bainbridge island	puget

Table 1 lists some of the novel relations learned by the proposed approach

Although the automatic approach can quickly evaluate large quantity of data, the reference set may not be complete or accurate. For sanity checking, we also randomly selected 30% of the relations marked incorrect by the automatic evaluator, which yields 203 relations of the North America induced by our approach. We then asked three judges (2 of them are graduate students; the other is an undergrad student in Computer Sciences) to label them correct, incorrect, or undecided. We found that 68 of them are marked correct by at least 2 evaluators. Some of such novel relations, listed in Table 1, include well-known facts, such as that the 'bronx' and 'coney island' are parts of 'new york city', and lesser known relations, such as that 'university of south florida' is in 'tampa'. Some relations would be judged incorrect by an expert, but in our opinion, they reflect the common "folk knowledge" shared by people. For example, most experts would not put 'pasadena', a city near La Cañada mentioned in the Introduction, in 'los angeles'. Lay people, on the other hand, often consider Pasadena as part of the "greater Los Angeles," and annotate their photos accordingly. As another example, 'disneyland' in not formally in the city of Los Angles ('la') or even the county of Los Angeles, yet this distinction is lost on many people who visit Los Angeles to go to Disneyland. The ability to discover such novel geospatial relations that reflect "folk knowledge" demonstrates the value of our approach.

In our proposed approach, most of the subsumption errors come from the place representation steps. Some places are very concave, even S-shaped. Think of Texas or Chile, for example. Our convex hull representation will inaccurately approximate these places by a convex region. Despite this, geographic subsumption approach significantly improves on the baseline method (Schmitz, 2006) for the following reasons. First, as observed by Schmitz (2006), users seldom annotate an image both with the most general and most specific tags. For example, using the baseline probabilistic subsumption method, p('university of south florida' | 'usa') = 0.0 and p('usa'| 'university of south florida') = 0.001. In other words, few users specify tags 'usa' and 'university of south florida' in the same photo. However, the geographic distribution of the tag 'usa' is likely to geographically subsume the distribution of the tag 'university of south florida'. Thus, geographic subsumption can solve the challenge of "general vs specific" concepts. Second, geographic subsumption can also solve the "popularity vs generality" challenge. For example, in baseline approach, p('california' | 'usa') = 0.14 and p('usa' | 'california') = 0.12. The result is that 'california' will subsume 'usa'. However, this is simply because users specify the tag 'california' more frequently than the tag 'usa'. Finally, with proper parameters, we can solve the "ambiguity" challenge, for example, Victoria in Canada or Australia or Cambridge in United States or United Kingdom. In fact, this challenge can also lead to the "popularity vs generality" challenge, because when evidence for ambiguous tags is aggregated, the total frequency may become more than its parent's. For example, the tag 'victoria' has been used 847,467 times on Flickr and 'british columbia' 513,116 times. However, 'victoria' tag could include instances of Victoria in Australia, and other places named 'victoria', resulting in a higher tag count for this concept than its parent 'british columbia' concept. After our method disambiguates the term 'victoria', it correctly infers that 'british columbia' geographically subsumes 'victoria'.

RELATED WORK

Several researchers have recently proposed approaches to learning conceptual hierarchies, or folksonomies, from social metadata. These approaches include graph-based (Mika, 2007), clustering (Brooks and Montanez, 2006) and hybrid methods that create similarity graph of tags (Heymann and Garcia-Molina, 2006). Schmitz (2006) has also applied a statistical subsumption model (Sanderson and Croft, 1999) to induce hierarchical relations of tags. All these methods use tags, and therefore, suffer from the "popularity vs generality" problem. Specifically, a certain tag may be used more frequently not only because it is more general, but because it is more popular among users. On *Flickr*, e.g., there are many more photos tagged with "Washington" than "United States". As was argued and demonstrated in the previous work (Plangprasopchok and Lerman, 2009), tag statistics alone may not be adequate for inducing relations.

In addition to tags, there are other types of user-generated metadata, such as set/collection hierarchies and geo-referencing tags (geotags) that are ubiquitous on the Social Web sites such. Geotags can potentially be used to resolve the "popularity vs generality" problem and many others by providing an additional view on how one concept geospatially relates to others. Researchers have begun to exploit geotags to induce "place semantics" – an association between place and other features, such as textual and visual information. Rattenbury and Naaman (2009) proposed an automatic approach to determine whether a certain tag is used for representing place(s). This approach is based on the assumption that, in general, a place tag appears locally, rather than ubiquitously, within a certain area. Meanwhile, a couple of recent works proposed frameworks to find correlations among geotags, visual features and tags of photos, and then utilize them for tag recommendation from photos" location (Moxley et al., 2008) and visual features (Kleban et al., 2009), or conversely, estimating location from visual features and tags (Crandall et al., 2009). The aims of these works are different from ours: we further investigate the approach to induce hierarchical relations among geospatial concepts to construct and/or enrich geospatial ontologies.

Several works dealt with the problem of disambiguating places. Approaches proposed by Li et al, 2003 and Amitay, et al., 2004 utilize gazetteers to identify places mentioned in some documents. In particular, when an ambiguous place is mentioned in a document, e.g., "Buffalo" can be one of 23 different cities in the United States, the rest of the document is scanned to obtain more clues, e.g. the term "NY". New clues in combination with the ambiguous name are then compared to some place names in the gazetteer. If there is one exact match, the place is then identified and hence disambiguated. Our method does not assume prior knowledge, such as a gazetteers, todisambiguate places, but locality of geographic coordinates and geographic subsumption relations. This way, our approach can enrich existing gazetteers, which, in turn, can be by other methods to achieve better performance.

CONCLUSION

We presented an approach for extracting geospatial knowledge from social metadata. We focus on diverse types of metadata on the photo-sharing site *Flickr*. We showed that we can use social metadata, such as sets and geotags, to discover and represent places and reason about them. The proposed method identifies specific instances of places and represents them geographically as convex regions. Geographic subsumption is then used to identify places that are parts of a given place. We showed that the method achieves reasonable performance on real-world data and is able to learn relations that do not exist in the reference set. The proposed method also improves on current state-of-the-art, probabilistic subsumption, which reasons about places based on the frequency of keywords describing those places. The improvements can be attributed to the proposed approach's better handling of the challenges associated with social tagging, for example, differentiating specific or popular vs general tags, also differentiating the senses of ambiguous tags. In future work we plan to improve place representation by using polygons and also incorporate evidence from collection/set relations.

ACKNOWLEDGMENTS

This work is supported, in part, by National Science Foundation under award CMMI-0753124.

REFERENCES

1. Amitay, E., Har'EI, N., Sivan, R., and Soffer, A. (2004) Web-awhere: geotagging web content. 27th Annual International
2. Brooks, C. H. and Montanez, N. (2006) Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proceedings of the International World Wide Web Conference*.
3. Crandall, D.J., Backstrom, L., Huttenlocher, D.P., and Kleinberg, J.M. (2009) Mapping the world's photos. In *Proceedings of the International World Wide Web Conference.*
4. Euzenat, J., and Shvaiko, P. (2007) *Ontology Matching*. Springer-Verlag
5. Golder, S. and Huberman, B. A. (2006) The structure of collaborative tagging systems. *Journal of Information Science*, Vol. 32, No. 2, pp. 198-208.
6. Graham, R.L. (1972) An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1: 132–133
7. Heymann, P. and Garcia-Molina, H. (2006) Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University.
8. Hopcroft, J.; Tarjan, R. (1973) Algorithm 447: efficient algorithms for graph manipulation. *Communications of the ACM*, 16: 372–378.
9. Kavouras, M., Kokla, M., and Tomai, E. (2006) Semantically-Aware Systems: Extraction of Geosemantics, Ontology Engineering, and Ontology Integration, *In Geographic Hypermedia (Lecture Notes in Geoinformation and Cartography)*, pp. 257-273.
10. Keating, T., and Montoya, A. (2005) Folksonomy Extends Geospatial Taxonomy. *Directions Magazine*.
11. Kleban, J., Moxley, M., Xu, J., and Manjunath, B.S. (2009), Global Annotation on Georeferenced Photographs, *In Proceedings of Conference on Image and Video Retrieval*.
12. Li, H., Srihari, R., Niu, C., Li, W. (2003) InfoXtract location normalization: A hybrid approach to geographic references in information extraction. In: Proceedings of the Workshop on the Analysis of Geographic References NAACL-HLT
13. Maths, A. (2004) Folksonomies: cooperative classification and communication through shared metadata. Technical Report, University of Illinois Urbana-Champaign.
14. Michalowski, M., Knoblock, C. A., Bayer, K.M., and Choueiry, B. Y. (2007) Exploiting Automatically Inferred Constraint Models for Building Identification in Satellite Imagery, *In proceeding of ACMGIS*.
15. Mika, P. (2007) Ontologies are us: A unified model of social networks and semantics. *Web Semantic.*, 5(1):5–15.
16. Moxley, E., Kleban, J., and Manjunath, B. S. (2008) Spirittagger: A geo-aware tag suggestion tool mined from flickr. In *Proceedings of MIR'08*.
17. Newsam, S. and Yang, Y. (2008) Integrating gazeteers and remote sensed imagery. *GIS*, 26.
18. Plangprasopchok, A. and Lerman, K. (2009) Constructing folksonomies from user-specified relations on flickr. In *Proceedings of the International World Wide Web Conference*.
19. Rashmi, S. (2005) A cognitive analysis of tagging. http://rashmisinha.com/2005/09/27/a-cognitive-analysis-of-tagging/
20. Rattenbury, T., and Naaman, M. (2009) Methods for extracting place semantics from Flickr tags. ACM Trans. Web., 3(1):1–30.
21. Sanderson, M. and Croft, B. (1999) Deriving concept hierarchies from text. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and Development in Information Retrieval, pages 206–213.
22. Schmitz, P. (2006) Inducing ontology from flickr tags. In *Proc. of the Collaborative Web Tagging Workshop (WWW '06).*