Discovering Points of Interest from Users' Map Annotations

Lakshmi Mummidi and John Krumm[1]

Microsoft Corporation

Abstract

One of the potential problems of volunteered geographic information (VGI) is ensuring its quality. Innocent mistakes and intentional falsehoods can reduce not only the quality of the information, but also people's confidence in VGI as a legitimate source of data. We present a case study in VGI that addresses the quality problem by aggregating input from many different people. Specifically, we present a technique to maintain a comprehensive list of points of interest (POI) for digital maps. This is traditionally difficult, because new POI are created, because some POI are known only locally, and because some POI have multiple names. We address this problem by exploiting map annotations contributed by regular, online map users. Our institution's mapping Web site allows users to create arbitrary collections of geographically anchored pushpins that are annotated with text. Our data mining solution finds geometric clusters of these pushpins and examines the pushpins' text and other features for likely POI names. For instance, if a given text phrase is mentioned frequently in a cluster, but infrequently elsewhere, this increases our confidence that this phrase names a POI. We tested the quality of our results by asking 100 local residents whether or not the POI we found were correct, and our user study told us we were

[1] Contact Author:
John Krumm
Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052 USA
jckrumm@microsoft.com
+1 (425) 703-8283 (phone)
+1 (425) 936-7329 (fax)

generally successful. We also show how we can use the same user-annotated pushpins to assess the popularity of existing POI, which is a guide for which ones to display on a map.

Keywords: points of interest, digital maps, online maps, data mining, volunteered geographic information

1 Introduction

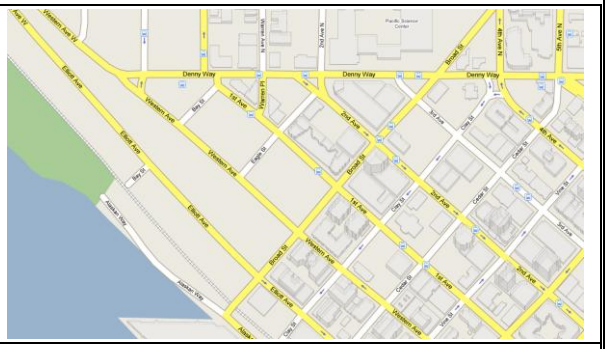
While volunteered geographic information (VGI) is a potentially attractive source of free data, its quality is not guaranteed. Manual vetting of the data can ensure quality, but this can be almost as expensive as generating the original data. We present a case study in VGI that addresses this problem by data mining a large collection of user-contributed geographic data, keeping only those parts that have sufficient support in terms of repeated contributions. Our algorithm sorts through a large volume of mostly irrelevant geographic data to find those parts that are mutually self-supporting. It then elevates these parts as valid geographic information. Specifically, we seek to exploit map annotations placed by regular users to find new points of interest (POI). In this way, our algorithm seeks to solve the problem of VGI data quality by processing a large volume of data in a way that is robust to the overwhelming proportion of useless data present. Techniques such as these are buoyed by the fact that online map usage is growing significantly, with comScore reporting a 33% growth in online map traffic in 2004 (http://blog.kelseygroup.com/index.php/2005/11/29/Online-Mapping-Outpaces-Overall-Internet-Growth/).

Online, consumer-focused maps typically display political subdivisions, street names, and natural features like rivers and lakes. They also endeavor to show certain points of interest such as parks, major institutions (*e.g.* universities and hospitals), museums, performance halls, businesses, and stadiums. Some online maps also show 3D models of buildings. The POI typically come from a database, which may be missing certain POI due to the relatively slow updates of the database. As an example, maps from the four major, consumer-level, online map providers are shown in Figure 1. These four maps show the region

surrounding a relatively new POI, "Olympic Sculpture Park", in Seattle, WA, USA. The new park is missing from all four, even at the highest zoom level, while nearby parks of about the same size are shown. This omission exemplifies the goal of our project: we want to find "Olympic Sculpture Park", along with many other missing POI, and add them to a map in the appropriate locations. We note here that our algorithm, described in subsequent sections, does find a "Sculpture Park" in the correct location, as shown in the figure.

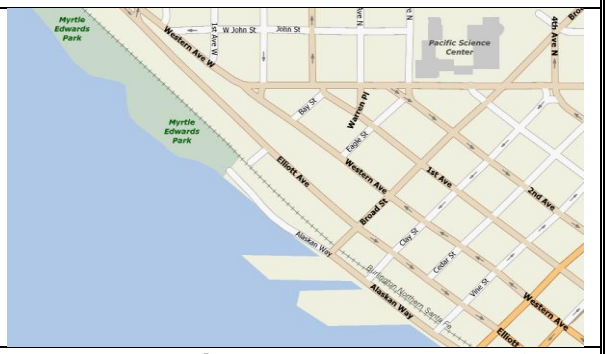
Michael Goodchild recently highlighted the potential use of "citizens as voluntary sensors" to create and enhance map data (Goodchild 2007). We endeavor to discover new POI from map annotations placed by regular users of our institution's mapping Web site, Microsoft's Live Search Maps (http://maps.live.com/). This site allows users to create collections of geographically anchored pushpins, as shown in Figure 2. Although a user's collection can be arbitrary, collections typically correspond to categories like restaurants, real estate, specialty shops, and many others. The site does not impose any taxonomy on the collections nor the pushpins, and we do not try to exploit any topical coherence among the pushpins in our POI-finding algorithm.

Microsoft Map (http://maps.live.com/)

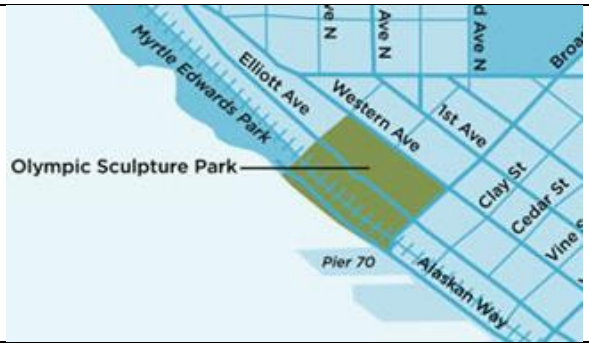


Google Map (http://maps.google.com/)

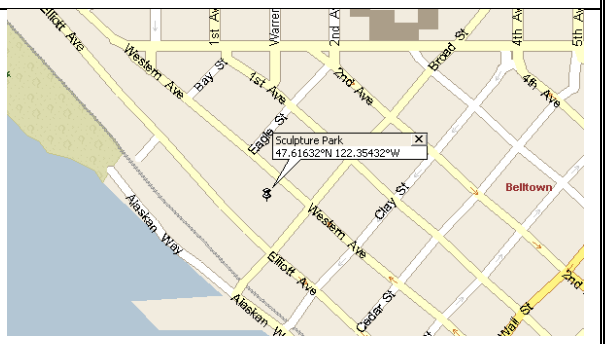
Yahoo! Map (http://maps.yahoo.com/)



MapQuest Map (http://www.mapquest.com/)



Map from sculpture park Web site



POI discovered by our algorithm

Figure 1: Four prominent mapping sites on the Web omit "Olympic Sculpture Park" in Seattle, WA as a point of interest. Based on data mining user contributions, our algorithm finds a "Sculpture Park" in the correct location.

The work most closely related to ours is World Explorer (Ahern, Naaman et al. 2007). The authors

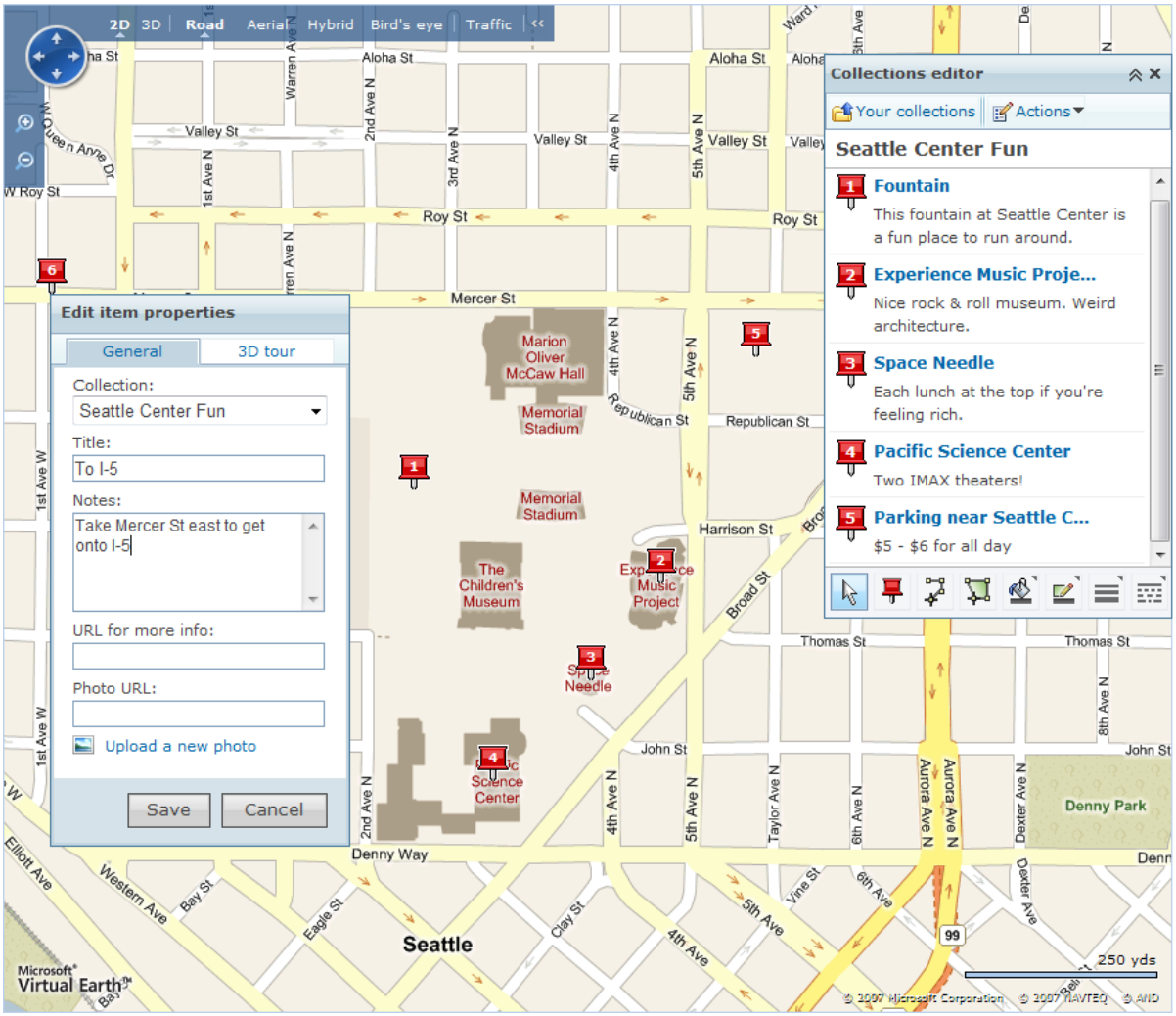


Figure 2: Users of our institution's mapping site can create arbitrary collections of pushpins. The panel on the right shows a collection, with a new pushpin about to be added from the panel on the left. Each pushpin has a name ("Title") and description ("Notes"), from which we extract candidate POI names.

processed captions on photos from Flickr, a photo-sharing Web site. Besides a user-supplied caption, each photo also had a user-supplied latitude/longitude associated with it. From these captions, they were able to extract meaningful place names, which they then attractively displayed on a map. Their algorithm starts with geometric clusters of the captions and candidate tag phrases extracted from the captions. World Explorer used TFIDF as part of their criteria for determining relevant phrases. TFIDF stands for the product of "term frequency" and "inverse document frequency" (e.g. (Salton and Buckley 1988)). For a candidate phrase in a cluster of captions, the term frequency is the number of times it appears in the

cluster's captions. Document frequency is the number of times the phrase appears in all the captions, both inside and outside the cluster. Dividing term frequency by document frequency gives TFIDF, and larger values of TFIDF indicate those phrases which appear frequently in the cluster and relatively infrequently outside the cluster. These are good candidates for a POI, because they are closer to unique for the cluster. We highlight the differences between World Explorer and our algorithm in the description of our algorithm below.

A related effort is WikiMapia (http://www.wikimapia.org/). WikiMapia lets users explicitly mark rectangles and polygons on a map to indicate places of interest. Users can edit, add to, and vote on places placed by previous users. Before a new place appears on the map, it must be approved by other users. Wikimapia is focused specifically on community creation of a place description layer. In contrast, the collections we mine are not necessarily intended for "clean" map labels, meaning we have to develop algorithms to take sometimes messy annotations and elevate them to reasonable POI names.

OpenStreetMap (http://www.openstreetmap.org/) is another related VGI effort in which users contribute map data. It is mostly aimed at creating road maps from users' GPS tracks or tracing roads on top of aerial imagery. Users can also edit contributed data for accuracy and cleanliness. As with WikiMapia, OpenStreetMap differs from our effort in that we use data that was not necessarily contributed for adding to a public map.

While mining user contributions for map additions is a recent idea, data mining for GIS is not. Examples include Li *et al.*'s (Li, Di et al. 2000) work on land use classification from data mining GIS sources. Anders (Anders 2001) gives a short taxonomy of data mining techniques for GIS and focuses on a graph-based clustering technique. Miller and Han's edited book (Miller and Han 2001) gives many additional examples.

Our technique exploits pushpins placed on an online map by regular users. We review this data in the next section.

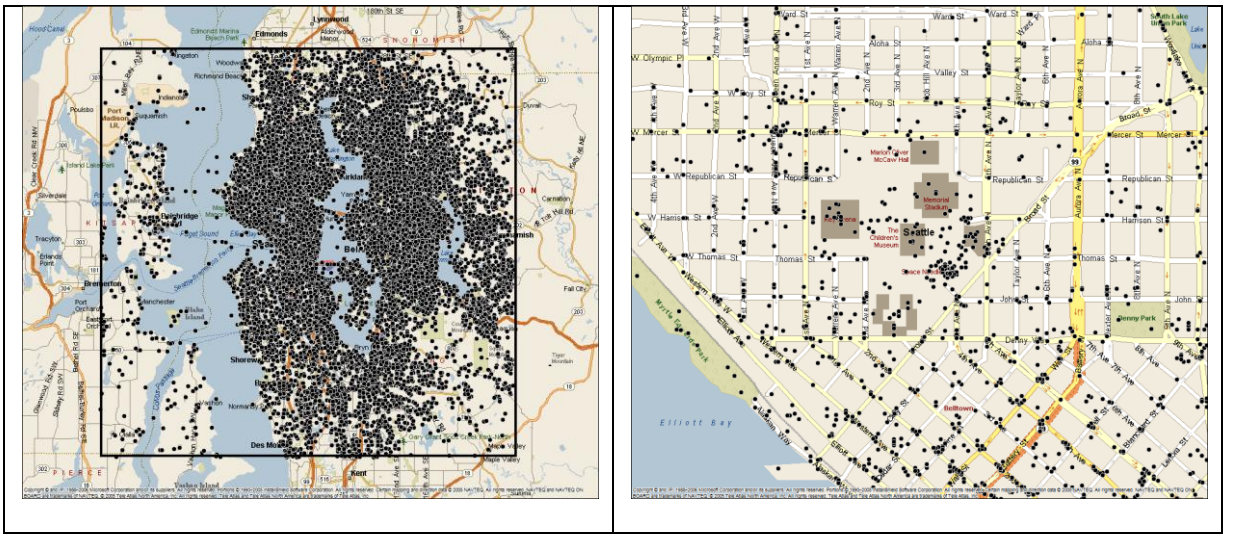


Figure 3: We analyzed 39,796 annotated pushpins inside the 45x45 kilometer square around Seattle, WA USA. On the right is a close-up of the map.

2 User Annotations on Pushpins

For purposes of experimentation, we limited our analysis to pushpins in the area surrounding Seattle, WA, USA. Figure 3 shows the locations of the 39,796 pushpin we used in a square about 45 kilometers on a side. This translates to 19.65 pushpins per square kilometer. The pushpins came from a snapshot of our collection database at the beginning of our experiments. Thousands of new collections are added every month. We limited our analysis to those collections that had been declared public by their authors. The pushpins we used came from 10,822 distinct collections, making an average of 3.68 pushpins per collection. We note that we used *all* the public pushpins in the bounded area, using no pre-filtering. Also, we ignore the fact that pushpins are grouped into collections, treating the aggregate of pushpins as one, large group to be mined for POI names. In a full production system, we would extract pushpins from all over the world, not just the region we chose for experimentation.

Each pushpin has a text field for a title and notes. These are entered by users as they add pushpins to their collection, and both fields are optional. Sample titles and notes from 50 pushpins inside our experimental square are shown in Table 1. 99% of the pushpins had titles, and 44% had notes. For those pushpins with titles, the average title length was 30 characters. For those pushpins with notes, the average

Table 1: These are the titles and notes of 50 pushpins from our experimental area. They were randomly chosen after eliminating pushpins with numerical digits in either of their text fields. Pushpins with numerical digits tend to be less interesting, such as street addresses. For this table, we also omitted pushpins with a missing title or note.

Title	Notes
Seattle, Washington	Veendam
Safeco Field	Seattle Mariners
Highlands Park and Ride	This is where Jill will drop you off and pick, you up.
First Stop in Seattle	This is in the International District
Bella Vista	Home Sweet Home
Taqueria Guaymas	A very good mexican restaurant on the East Side
Mayflower Park Hotel	Estancia en el centro de Seattle downtown. Justo frente a la
Cherry Street Coffee House	Kindra please help
Garage Billiards	The Garage is a pool hall and bowling alley where you can
The McLeod Residence	The McLeod Residence is an art gallery/hip hangout in
Highland Ice	Highland Ice Arena
Soccer Field Parking Entrance	This is the best entrance to use to park for the soccer field.
Main Entrance	Use this entrance to park for the soccer fields.
Port-a-Potty	There is a bathroom here!
Parking Lot Entrance for Robinswood Park Fields	This is the entrance to the parking lot for the Robinswood Park
Lakeridge Elementary Fields	This is the soccer fields area for Lakeridge Elementary School.
Parking Lot	This is the Lakeridge Elementary parking lot.
Parking Lot - South Mercer Playfields	This is the parking lot for the South Mercer Playfields.
Play Structure	South Mercer Playfields play structure
Soccer Field	This are is used for a full size soccer field.
Soccer Field	This area is used for a full size soccer field.
Soccer Field	This area is used for a youth soccer field.
Soccer Field	This area is used for a youth soccer field.
Softball Field	South Mercer Playfields have four adult softball fields.
Lakeridge Elementary School	Arrive at Lakeridge Elementary School.
East Entrance	This is the entrance to the parking lot.
Route to Parking Lot	This is the route to the parking lot from the east entrance. It is
West Entrance	This is the west entrance to the parking lot.
Eastside Youth Soccer Association (EYSA)	Bellevue, WA
Highline Soccer Association (HSA)	Burien, WA
Lake Washington Youth Soccer Association (LWYSA)	Redmond, WA
Seattle Youth Soccer Association (SYSA)	Seattle, WA
Entrance	Use this pushpin for directions to and from the church.
Entrance	Use this push pin for directions to and from the center.
Entrance to Parking Lot	Use this pushpin for directions to and from the stadium.
Entrance	Use this pushpin for directions to and from Soccer Nation.
Entrance	Use this pushpin to get directions to/from the fields.
Microsoft Samm E Building	Sammamish Campust
Parking	Albert Einsten Elementary School Parking
Turner HMC Parking	Enter parking lot from Jefferson Street.
Fireworks	Myrtle Edwards Park (park), Seattle, Washington, United
Chengdu	Chinese buffet
Pacific Science Center	IMAX and science center
Space Needle	May want to go to the top
Discovery Park	Discovery Park (city park), Seattle, Washington, United States
Elliott Bay ViewPoint	Harbor Ave SW
Light Rail Station Beacon Hill	Beacon Avenue South and South Lander Street
Kirklands bibliotek	Ett st??rre bibliotek ligger i Bellevue men hit g??r barnen
PCC	H??r handlar vi den mesta maten (ekologiskt odlat!)
Starbucks Coffee	Denna om??ttligt popul??ra kaffekedja finns numera v??rlden

note length was 64 characters. All pushpins have a latitude/longitude pair, which is determined by the user clicking on the map.

3 From Pushpins to Ngrams and Clusters

After extracting the pushpins' locations, titles, and notes, we have a set of text annotations anchored to locations specified by latitude/longitude. The next step of our algorithm is to extract candidate POI names from the text along with numerical features with which to assess their suitability for adding to a map. For instance, we see many unsuitable candidate phrases (e.g. "area is used"), so we need to have a way to filter out these garbage phrases. We begin by extracting candidate phrases, called ngrams, and clustering pushpins to find groups that are physically compact. We then compute numerical features of the ngrams inside each cluster with an eye toward find suitable POI names. This section discusses extraction of candidate ngrams and clustering.

3.1 Ngrams

From each pushpin, we extract candidate POI phrases from its title and notes. Specifically, we extract ngrams for N = 1, 2, and 3: monograms, bigrams, and trigrams, respectively. (An ngram is a phrase with n words in it.) For instance, the description of one of the pushpins we have is "South Mercer Playfields play structure". From this, we would extract the 12 ngrams in Table 2.

Table 2: These are the 12 ngrams (monograms, bigrams, and trigrams) we would extract from "South Mercer Playfields play structure".

South	Mercer	Playfields	play	structure	South Mercer
Mercer Playfields	Playfields play	play structure	South Mercer Playfields	Mercer Playfields play	Playfields play structure

Except for computational speed, there is no reason not to consider ngrams with more than three words. We stopped at three so we could carry out our experiments in a reasonable amount of time.

Note that the extracted ngrams come from adjacent words in the original text, so we do not construct ngrams that skip over words. For comparing ngrams, we ignore the case (upper or lower) of the

Table 3: We eliminated ngrams that contained any of these words to help eliminate non-interesting phrases. These are mostly "stopwords" that tend to join separate phrases, but also words that we found frequently as part of phrases that were clearly not related to POI. We also eliminated two-letter abbreviations for U.S. states and numerical digits.

Stopwords					Streets	Compass Directions
a	down	in	take	we	av	n
about	etc	is	than	went	ave	ne
after	even	it	that	what	blvd	e
also	every	my	the	when	circle	se
although	for	new	their	where	court	s
an	from	nice	them	who	cr	sw
and	get	no	then	will	ct	w
any	go	not	there	with	lane	nw
are	good	now	they	www	ln	n
as	had	of	this	year	st	ne
at	has	on	though	yes		e
be	have	or	to	you		
better	her	our	too	your		
between	here	out	took			
but	him	over	type			
by	his	part	up			
can	home	quite	very			
com	how	see	want			
could	hr	select	was			
day	i	so	way			

characters. We split phrases into words at each instance of one or more adjacent space characters. We ignore punctuation marks except for apostrophes, which can be a legitimate part of a POI name. We also ignore ngrams with so-called "stopwords" and other words listed in Table 3. This helps eliminate garbage ngrams that likely do not name a point of interest, and it reduces processing time.

After processing our 39,796 pushpins, we found 9,674 distinct monograms, 20,403 distinct bigrams, and 9,005 distinct trigrams for a total of 39,082 distinct ngrams. With an area of 45^2 kilometers, this translates to 19.30 distinct ngrams per square kilometer.

Our intuition is that the annotations in clusters of nearby pushpins will contain a suitable name for that part of the world. The ngrams described above are candidate names. The next section describes how we form candidate clusters.

3.2 Clustering Pushpins

In creating candidate ngrams, we cast a wide net and extracted all possible ngrams, for N = 1, 2, and 3, from the text annotations, minus some predefined words and characters to reduce the overall number. In creating candidate clusters, we follow a similar philosophy in that we generate many more clusters than we ultimately use. This approach helps ensure we preserve most of the good candidates.

The goal of clustering is to find groups of nearby pushpins. We chose to use a dendrogram (see (Duda and Hart 1973)) for this purpose. The dendrogram manifests a hierarchical agglomerative clustering technique. At the beginning, each pushpin is its own cluster. Each subsequent step merges the two clusters that are nearest to each other, based on their latitude/longitude. When two clusters merge, their new location is taken as the centroid of their constituent pushpins. At the highest level, all the pushpins are in the same cluster. The dendrogram does not give a good indication of the optimal number of clusters, *i.e.* when to stop merging. We consider all the possible clusters when we look for new POI names.

Clustering pushpins with a dendrogram can be computationally slow, because it requires a distance computation between all unique pushpin pairs. We reduced the computational time by splitting our 45x45 kilometer test region into 4x4 equally sized, square subregions. We computed a separate dendrogram for each subregion. While this risks splitting small clusters that span the boundary between two subregions, we were not worried about eliminating large, spanning clusters, because most good POI names come from localized groups of pushpins.

3.3 Ngram/Cluster Parameters

Each cluster of pushpins normally has many ngrams. Also, the same ngram can appear in more than one cluster. We process distinct ngram/cluster pairs to find those ngrams that seem appropriate for adding to the map. After clustering, we found over 1.7 million ngram/cluster pairs, from which we need to extract the ones with relevant POI names. Toward this end, for each ngram/cluster, we compute a few numerical

parameters which we use to assess whether or not the ngram represents a good point of interest. Before describing these parameters, we note that each cluster has a latitude/longitude centroid computed from its constituent pushpins. This centroid is where we would place the ngram on the map if it is elevated to a POI.

One of the most important ngram/cluster parameters is "term frequency inverse document frequency" (TFIDF) (Salton and Buckley 1988), described above in the context of World Explorer. TFIDF is commonly used in document search and retrieval applications. In our application, one ngram/cluster serves as a single document. "Term frequency" (TF) is the number of pushpins in the cluster that contain the ngram. A high TF could be evidence that the ngram is significant and should be extracted as a POI name. "Document frequency" (DF) measures how often the ngram appears in all the pushpins, including those outside the cluster. A high DF indicates that the ngram is not specific to the cluster in question. TFIDF is TF/DF, which is indicative of the ngram's frequency inside the cluster and infrequency outside the cluster. A high TFIDF is evidence for a good POI name. As an example, the monogram "park" might occur very frequently in many clusters, giving it a high DF. Thus, anytime "park" comes up as a candidate ngram in a cluster, its TFIDF will be low. In fact, the list of stopwords in Table 3 are some those we expect will have a high DF, so they are eliminated even before consideration.

World Explorer (Ahern, Naaman et al. 2007) uses TFIDF as part of their criteria for selecting POI names from geotagged Flickr photos. They noticed, however, that TFIDF could be artificially high if a single photographer used the same text tag for a larger number of photos. This raises the TF of a cluster, but such tags were found to be not necessarily descriptive. They avoid such clusters by computing the fraction of photographers in each cluster who used the tag in question. A higher fraction indicates that more photographers considered the tag relevant. They threshold the product of TFIDF and this fraction to decide which captions to extract as POI.

Besides TFIDF, we use two other parameters to find good POI names. As described above, our dendorgam clustering procedure grows clusters of pushpins without an upper size limit. In fact, the clusters keep growing until all the pushpins in each subregion are grouped into one, large cluster. Such a large cluster is likely not indicative of a single POI, so we compute a parameter that tends to identify clusters where a large fraction of the pushpins mention the POI in question. We call this parameter "term purity", and we compute it as the fraction of pushpins in the cluster that contain the ngram.

The final parameter we use is simply the number of pushpins in the cluster. A lower bound on the number of pushpins tends to eliminate garbage phrases. As an example, a single pushpin with a single, unique (and possibly strange) ngram qualifies as a cluster, has the maximum possible TFIDF (1.0) and the maximum possible term purity (1.0). The threshold on the number of pushpins helps eliminate the strange ngrams that come from pushpins like this.

The three ngram/cluster parameters we use are summarized in Table 4. For all three of these parameters, a larger value is better, so we apply a lower threshold to find good POI names. We discuss the extracted POI and evaluate the method in the next section.

Table 4: These are the parameters that we used to identify good clusters and their associated ngrams for POI.

Parameter	Meaning	Range	Lower Threshold
TFIDF	"term frequency, inverse document frequency" assesses the distinctiveness of an ngram inside a cluster compared to everywhere outside the cluster	0.0 – 1.0	0.8
term purity	fraction of pushpins in cluster that contain ngram	0.0 – 1.0	0.8
number of pushpins	number of pushpins in cluster	≥ 1	5

4 Assessment of Extracted POI Names

With the three parameters described in the previous section, we did a small amount of tuning to find a good set of lower thresholds to apply in order to extract good POI names. We settled on the thresholds shown in Table 4. The nature of our hierarchical clustering means that the same ngram will appear in many different clusters. After thresholding, for ngram/clusters with the same ngram, we picked the ngram/cluster with the most pushpins. This thresholding and duplicate elimination resulted in 286 distinct POI names. The first 100 unedited POI names we found are shown in Table 5. The corresponding locations of the POI were computed as the centroid of the locations of the pushpins in the cluster.

Looking at these POI names, we see some spelling mistakes and unconventional capitalization, meaning they would probably require some editing before being added to an official POI database. In their unedited form, these POI could be presented to a user as what they are: casual contributions from

Table 5: These are 100 of the 286 POI we found. We have not edited the ngrams.

Trails Apartments	Off Broadway	residential renovation	Cured Meats	Town Center Cinema
Robinswood Park Bellevue	Anne High School	Chili Parlor	X HEADQUARTERS	Macrina Bakery Cafe
Stellar Pizza	Pierre Ford	Glazers	Steps Apartments	Wine Room
corner Columbia	Bay Book Co	under carpets	Foods Bellevue	Hjarta
Ivanhoe East	Emylie Witte	Morgan Junction	Seatac Park	Great ZAA!
Swim Bouy Line	States Duma	Foods Persian	American School-Puget	Bouy Line
Roadhouse Casino	LEED certification	Science Labratory	Round Baby News	Valley Park Issaquah
Palace India Cuisine	PERSON GALLERY	rocket ship	Oyster Lng	running loop
Cyclo Cafe	Creations Ltd	Finn Hill Park	cash prizes	Inglemoor High School
Mill Burgers	Redmond,Washington United States	Field Seattle Mariners	pont du ship-canal	Digipen
Field/King	Mosler	Island Presbyterian Church	Pelican Cafe	Roger Taproom
Clipper Reservations	Cycle Inc	Bag Cafe	Lauren Catalano	Gilman Playground
Camera Supply	Sunday Farmers Market	Bay Brewery Pub	feta cheese	Mary Hoey
Ales Brewery	Pacific Regent	Gary Schimek	Bridget Johnston	Greenlake Cycle Inc
Piper Ale House	Saucer Pizza	Kenzie Smernis	Check-Out: TERRY HALL	Vegetarian House
MacRina	Byrne Pub	Brands Floor Supply	Inn Roadhouse Casino	Lawton Training Area
Microsoft LZ	Learning Lab	Park Bellevue	Northwest School-Music	Marriott Kirkland
CLOUD INN EASTGATE	Eric Luksetich	Hoa Restaurant	Artisan Cured Meats	Garden Chinese Restrnt
Vegetarian Place	bar convenience store	Leota	Cue Billiards	Rikki
Hosoonvi	Film Forum	outdoor fire	Center Cinema	Hefeweizen

other users that have been deemed somehow significant enough to show on a map.

Other forms of volunteered geographic information can suffer from similar quality issues. Volunteers may not know nor care about the quality of their contributions. In our particular case, users do not necessarily intend to have their pushpins shown to the public, so they may not be concerned about spelling or positional accuracy. In addition, contributors can easily maintain anonymity by creating a new, free, Web-based email account specifically for their collections. Thus, there is no risk in creating mistaken pushpins. An interesting research study in VGI would be to assess the quality of contributed data as a function of the contributor's anonymity and their perception of how the data will be used.

We carried out a small user study to assess the correctness of the POI we found in order to see if our algorithm really is good enough to extract useful information. Our user study consisted of showing a map with some of our POI superimposed as labels. An example map is shown in Figure 4. We made 100 maps with POI randomly chosen from those we found, and gave each map to one of 100 different employees at our institution. These employees all live in the same area covered by our map, so they have some

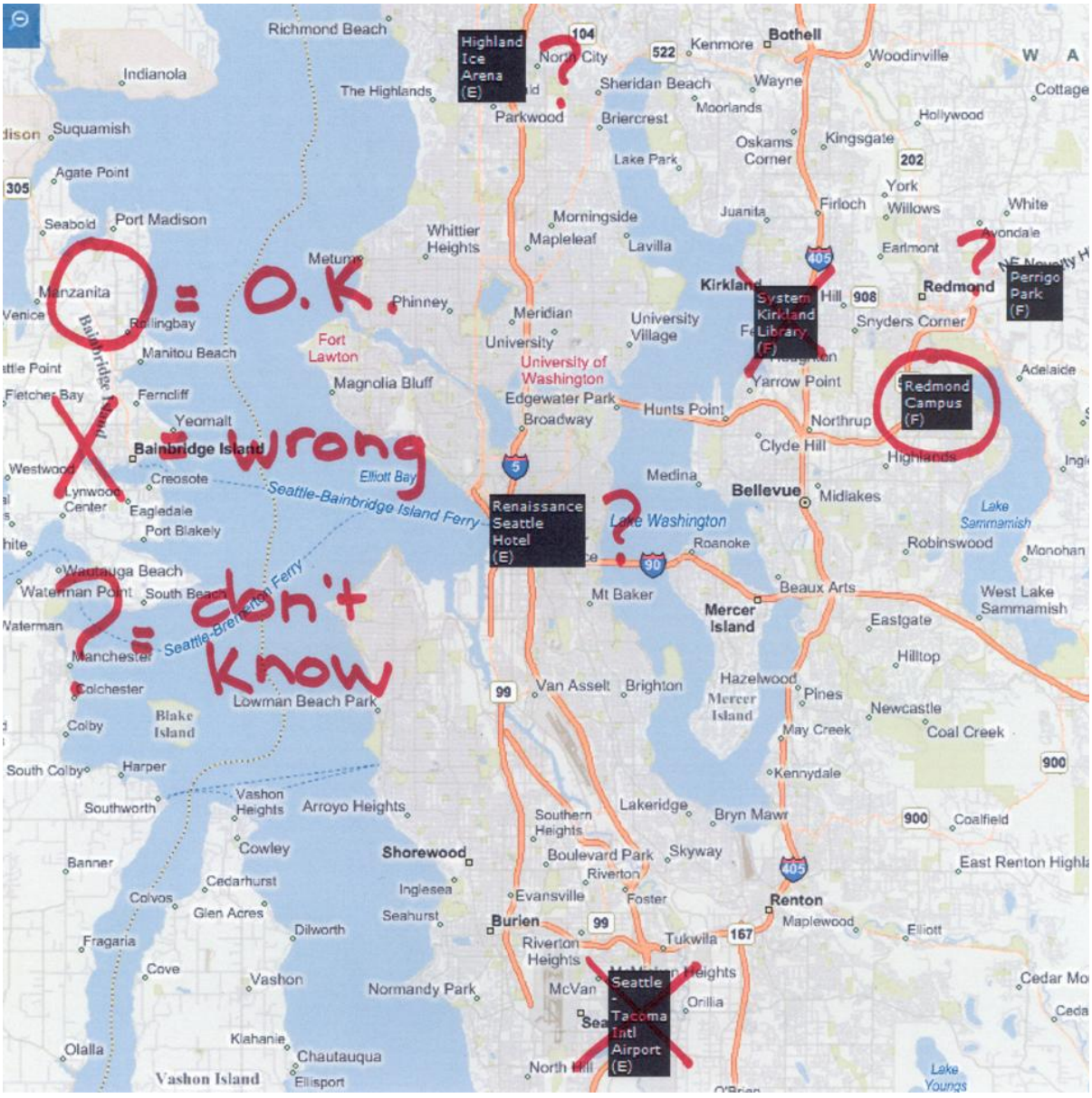


Figure 4: We asked users to indicate what they thought of the POI we extracted by marking a paper map. The "(E)" and "(F)" stand for "existing" and "found " POI, although we did not make this distinction to our users.

familiarity with the area. For each POI superimposed on the map, we asked the subject to mark it in one of three ways:

-- Circle the POI if it was recognized and in the approximately correct location
-- Cross out the POI if it was recognized but misnamed or in the wrong location
-- Add a question mark if the POI was not recognized

As a control, we split the POI we found into two categories. One category, "existing POI", were ngrams from our list that we also found in a Yellow Pages database of businesses. The other category, "found POI", were the ones that were not in the database. With this distinction, we could assess how users rate the new POI we found against existing POI already in a database. We note, however, that even the existing POI came from our data mining process, so they are not necessarily in the correct location, and they may have unconventional capitalization. Each of the 100 maps had three existing and three found POI on it. Sometimes the labels overlapped enough to obscure one, and we told our users to ignore the label if they could not read it in these cases.

Table 6: These are the results of our user study. The proportions of unrecognized, correct, and wrong POI was about the same for both existing and found POI.

	Unrecognized	Correct	Wrong
Existing POI	82.4%	16.2%	1.4%
Found POI	80.1%	15.3%	4.6%

Table 6 gives the results of our user study based on 100 users viewing an average of 5.72 POI on the map. The resulting proportions for the two conditions, "existing" and "found" are fairly close to each other. By far the most frequent response for both conditions was "don't know", at 81.3% of the total for both categories. This indicates that the POI we asked about were generally not recognized.

Limiting the scope of the analysis to only those POI that were recognized, existing POI were deemed correct 92.2% of the time, while found POI were deemed correct 76.8% of the time. Thus, existing POI

were correct more often, but found POI were correct a significant amount of the time, meaning that our algorithm is somewhat successful in extracting meaningful labels to add to the map. We note that existing POI were not always considered correct. This may be due to unfamiliarity on the part of our subjects or out-of-date Yellow Pages.

This is the first objective assessment of volunteered map annotations that we know of.

The advantage of this technique is its large and growing source of data. The existing mapping site is enough of an incentive to attract millions of pushpins, and the number is growing. The growth means that the data is able to keep up with new points of interest as they come into existence. The incremental cost of the refined data is very low, as our algorithm can run on top of the existing infrastructure already created for gathering user collections.

Some challenges regarding this technique include the question of aging out old data that may represent obsolete, renamed, or repositioned POI. Also, as the system stands, there is very little incentive for correct spelling or placement of pushpins, leading to many mistakes in the data that our algorithm has to work around. We have not yet explored how well the algorithm works with a reduced number of pushpins nor in regions with a mix of languages.

5 POI Importance

Another simpler use of our pushpin data is to assess the popularity of known POI. Fundamentally, a pushpin casts a vote each time it mentions an existing POI. This is potentially useful for deciding which POI to display on a map that has only limited space or resolution. It would also be useful for browsing maps to see what users consider the most interesting places to know about.

In order to test this idea, we queried our database of user-supplied pushpins to find bigrams and trigrams that we could also find in our Yellow Pages database, guaranteeing that the ngrams actually

represented existing POI. Due to partial matches to Yellow Pages entries, matched monograms were mostly garbage. Each pushpin that mentioned one of the ngrams counted as one vote.

Sometimes, high vote-getters represented chain stores, which are distributed over the region. We eliminated these by computing a geometric spread for each ngram. Specifically, we computed the "median absolute deviation" (MAD) (Rousseeuw and Croux 1993) of the voting pushpins' latitude and longitude. The MAD is a robust estimate of a scalar's variation. We converted the MAD of the latitude and longitude for each ngram into meters, took the maximum of these two values, and eliminated any ngram whose maximum MAD was greater than one kilometer. This helped ensure that the resulting POI were compact.

The top 30 compact POI from our study are shown in Figure 5 along with the number of votes for each one. These are generally familiar to area residents and include prominent tourist attractions (Space

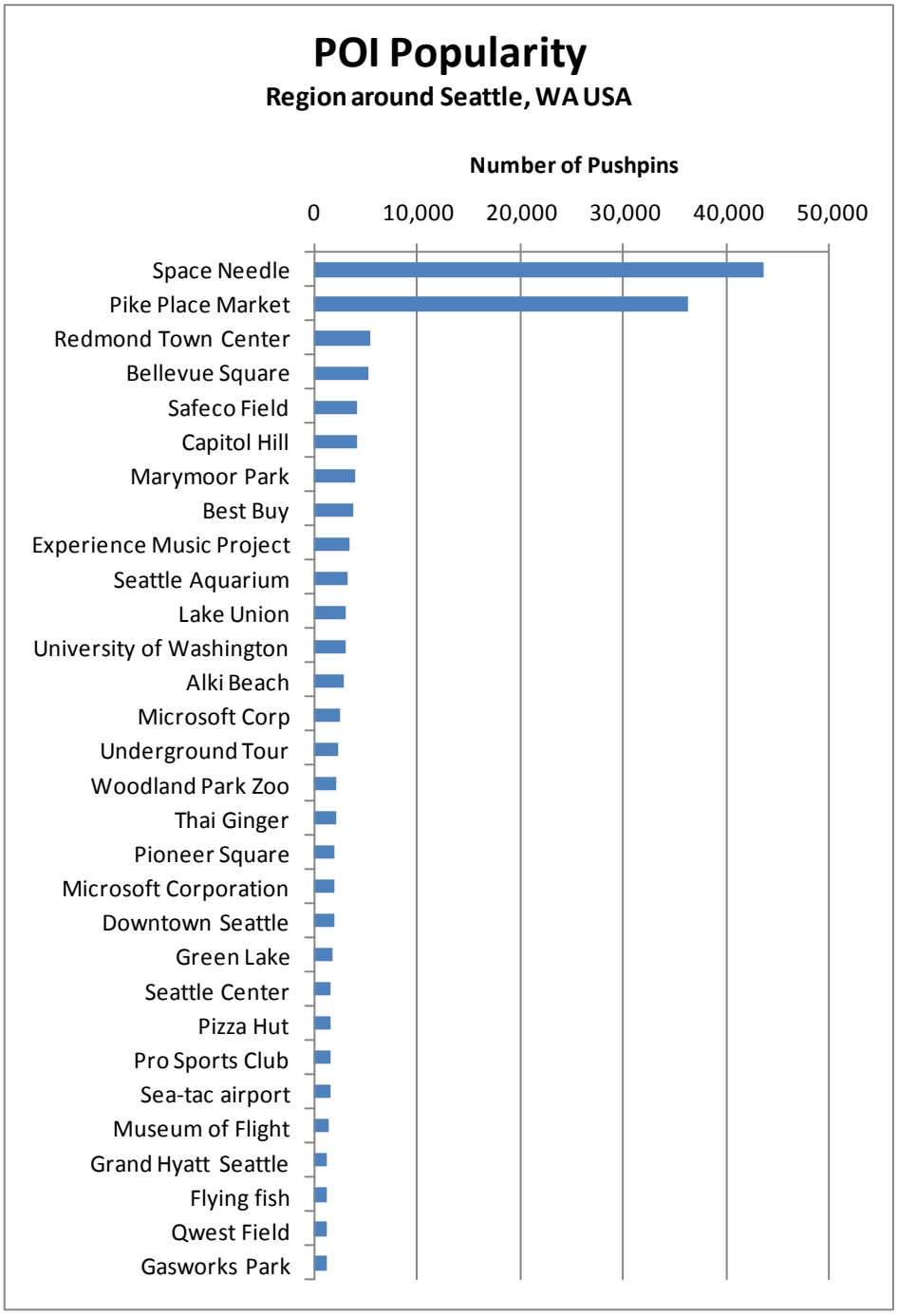


Figure 5: These are the 30 most popular points of interest in the Seattle, WA USA area based on votes from user-supplied pushpins.

Needle, Pike Place Market, Experience Music Project, Seattle Aquarium), malls (Redmond Town Center, Bellevue Square), sports arenas (Safeco Field, Qwest Field), and major institutions (University of Washington, Microsoft Corporation). Despite our insistence on compactness, the list also includes a few chain stores (Best Buy, Pizza Hut).

6 Conclusion

This paper seeks one way of ensuring the quality of volunteered geographic information. Instead of an expensive process of manually verifying VGI, our algorithm searches through a large collection of VGI and retains only those parts that are consistently repeated. Specifically, this paper describes an algorithm for finding new points of interest to add to a map based on user annotations. We examined almost 40,000 annotated pushpins placed around the Seattle, WA USA region by regular users of a mapping Web site. We extracted monograms, bigrams, and trigrams as candidate POI names from the text associated with each pushpin. After clustering the pushpins by location, we found reasonable POI names based on simple features of the clusters. We categorized the POI we found as either existing or found, depending on whether or not the POI exists in a Yellow Pages database, respectfully. Our user study showed that of the POI that were recognized by users, existing POI were deemed correct 92.2% of the time, and found POI were deemed correct 76.8% of the time.

Not all the POI we found were correct, and some of them had slight mistakes with capitalization and spelling. Thus, the points of interest that we extracted are likely not suitable for adding to a base map without some editing. Alternatively, they could be added to a layer specifically designated as user-contributed in order to correctly set expectations. This is what World Exploer does (Ahern, Naaman et al. 2007).

We also showed, using much simpler processing, how to assess the popularity of POI in an existing database based on votes from user-supplied pushpins. This could be used to decide which POI to show on a map or as a popularity layer for map browsing.

We envision a few possible extensions to this research. One interesting problem would be to put found POI into taxonomy (e.g. restaurant, grocery store, museum, etc.) automatically based on text associated with the pushpin cluster, including possibly the title and description of the pushpins' original collection. Another avenue to explore is to automatically determine the extent of a point of interest such as a lake or park. We place the POI at a cluster's centroid, but the extent of the cluster might be a good indication of the POI's extent. Finally, since new pushpin collections are being added at the rate of thousands per month, an analysis similar to ours could be used to find trends in the popularity of POI. For instance, we expect major events like the Olympics temporarily shift people's focus to different parts of the world.

References

Ahern, S., M. Naaman, et al. (2007). World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. Seventh ACM/IEEE-CS Joint Conference on Digital Libraries, (JCDL 07). Vancouver, Canada.

Anders, K.-H. (2001). Data Mining for Automated GIS Data Collection Photogrammetric Week 01. Heidelberg, Germany: 263-272.

Duda, R. O. and P. E. Hart (1973). Pattern Classification and Scene Analysis, John Wiley & Sons.

Goodchild, M. F. (2007). "Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0." International Journal of Spatial Data Infrastructures Research 2: 24-32.

http://blog.kelseygroup.com/index.php/2005/11/29/Online-Mapping-Outpaces-Overall-Internet-Growth/.

http://maps.live.com/.

http://www.openstreetmap.org/.

http://www.wikimapia.org/.

Li, D., K. Di, et al. (2000). "Land Use Classification of Remote Sensing Image with GIS Data Based on Spatial Data Mining Techniques." International Archives of Photogrammetry and Remote Sensing 33(B3): 238-245.

Miller, H. J. and J. Han, Eds. (2001). Geographic Data Mining and Knowledge Discovery, Taylor & Francis

Rousseeuw, P. J. and C. Croux (1993). "Alternatives to the Median Absolute Deviation." Journal of the Americal Statistical Association 88(424): 1273-1283.

Salton, G. and C. Buckley (1988). "Term-Weighting Approaches in Automatic Text Retrieval." Information Processing & Management 24(5): 513-523.