Balancing Diversity to Counter-measure Geographical Centralization in Microblogging Platforms

Eduardo Graells-Garrido
Universitat Pompeu Fabra
Barcelona, Spain
eduard.graells@upf.edu

Mounia Lalmas
Yahoo Labs
London, UK
mounia@acm.org

ABSTRACT

We study whether geographical centralization is reflected in the virtual population of microblogging platforms. A consequence of centralization is the decreased visibility and findability of content from less central locations. We propose to counteract geographical centralization in microblogging timelines by promoting geographical diversity through: *1)* a characterization of imbalance in location interaction centralization over a graph of geographical interactions from user generated content; *2)* geolocation of microposts using imbalance-aware content features in text classifiers, and evaluation of those classifiers according to their diversity and accuracy; *3)* definition of a two-step information filtering algorithm to ensure diversity in summary timelines of events. We study our proposal through an analysis of a dataset of Twitter in Chile, in the context of the 2012 municipal political elections.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

Keywords

Information Filtering; Information Diversity; Geolocation.

1. INTRODUCTION

Microblogging sites are social platforms where, worldwide, users are able to participate; there are no physical barriers. However, geography still plays an important role in the way user content is generated [1, 8, 18, 19, 25], not always in a fair manner. When population distribution is imbalanced, it is expected that content shared on social platforms is also imbalanced, but in centralized populations, content from non-central locations may itself be driven by the central locations, further increasing the imbalance. This causes two problems: on one hand, the voice of less populated and non-central locations is lost in a timeline flooded with content from few locations. On the other hand, algorithms to classify content become biased because of the over-representation of content generated from centralized places, particularly when algorithms rely on representative documents for modeling. For instance, predicting that all content is about the centralized and most populated location will lead to high accuracy, in spite of the absence of geographical diversity.

In Chile, around 40% of its population live in its capital, Santiago, and public policy and media are biased towards the needs of the capital [23]. A common saying is *"Santiago is not Chile"*, referring to the fact that the capital is not representative of the country, yet media outlets concentrate in Santiago and government policies are tailored towards its needs. Given the climatic, geographical and cultural diversity of Chile, centralization is a serious problem.

Motivated by the above situation, we address two research questions: **Q1)** *When centralization is present in the physical world, is it reflected in the virtual population of microblogging platforms?* **Q2)** *How to generate geographically diverse event-specific timelines?* We study both questions through a case study using a dataset from Chile focusing on the microblogging platform Twitter. On October 28, 2012, we crawled microposts, or *tweets*, published in the context of municipal elections held in Chile that day [24]. This event was locally relevant throughout the country, making it a good dataset for our research questions as local events have denser discussion networks than global events [25].

Many tweets are not associated with a geographical location, so we need to geolocate their content to provide geographical meta-data. To this end, we geolocate users by querying a high precision gazetteer [10, 15, 22]. Then, to geolocate tweets we extend previous work [9] by considering content features aware of user locations. These features are based on TF-IDF (which has superior performance than topic modeling for recommendation tasks [20]) and allow us to train classifiers with a number of features that are several orders of magnitude smaller than typical bag of words approaches. To overcome the false sense of accuracy introduced by imbalance, we consider both accuracy *and* diversity when evaluating, because *accuracy is not enough* [13]. Then, building on prior work [7, 16] we define an information filtering algorithm to generate a diverse timeline with a focus on geographical diversity that summarizes an event.

Our work makes the following contributions: *1)* we show that centralization from the physical world is reflected in a spatially representative sample of the Chilean virtual population in Twitter; *2)* we address population imbalance from a content perspective, while previous work has focused

HT'14, September 1–4, 2014, Santiago, Chile.

ACM 978-1-4503-2954-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2631775.2631823.

on a network perspective [22]; *3)* to evaluate classifiers considering accuracy and diversity, we define a *D-measure* based on the F-measure used in Information Retrieval [2]; and *4)* we evaluate our information filtering algorithm, and find that timelines generated with our approach ensure geographical diversity, and are more diverse than plain timelines based on popularity.

2. METHODOLOGY

We focus on timelines from microblogging platforms. Even though our definitions are general, we restrict ourselves to Twitter. Twitter is a microblogging platform where users publish status updates called *tweets* with a maximum length of 140 characters. Users can *follow* other users, establishing directed connections between pairs of users. When user A *follows* user B, tweets and *re-tweets* made by B will show up in A's timeline. A timeline is a list of tweets in reverse chronological order. Users can annotate tweets using *hashtags*, i.e. keywords that start with the hash character #.

Problem Definition. We define our problem as follows: *1)* given an event E (defined as a set of hashtags and special keywords) relevant to a country C, with a set of locations L, collect all tweets related to E generated in C in a tweet set T_E; *2)* given all users U who published tweets in T_E, predict (if possible) a location from L for all users $u \in U$; *3)* considering the users U_L who were geolocated, aggregate their interactions to find if geographical centralization is present; *4)* aggregate the content from U_L into location documents, and use these to build a location classifier P such that given an arbitrary tweet, predicts its location; *5)* using the output from P applied to all tweets in T_E, filter T_E to produce a summary tweet set T_θ, with $|T_\theta| \leq |T_E|$, which is more geographically diverse, i.e. $geodiversity(T_\theta) \geq geodiversity(T_E)$.

Geolocating Users. To geolocate users we rely on the self-reported location in user profiles. Instead of querying external services using profile locations as input [15], we build an ad-hoc gazetteer from official location names, from lists of known toponyms extracted from Wikipedia [10, 22] and from labeled user profiles. Then, to geolocate a user u, we query the gazetteer with u's self-reported location.

Location Interactions. In previous work, two locations become connected if someone from location A follows someone from location B [12]. In our context this is not meaningful, because such connectivity may not convey an interaction between two users that is relevant to the event E. Hence, we consider *1-way interactions* between locations through mentions and retweets [19] by building an adjacency matrix: $M_{i,j} = mentions(L_i, L_j) + retweets(L_i, L_j)$, where $mentions(L_i, L_j)$ is the number of tweets from location L_i (those tweets whose author has been geolocated to that location) that mention one or more accounts from location L_j, and $retweets(L_i, L_j)$ is the number of times that tweets from L_j have been retweeted by users from L_i.

Measuring Centralization. From the adjacency matrix we build an undirected graph with self-connecting edges removed, and estimate the edge weights with a normalized *geometric mean* of information flow:

$$weight(i,j) = \frac{\sqrt{M_{i,j} \times M_{j,i}}}{\max\{\sqrt{M_{i,j} \times M_{j,i}} | \forall l_i, l_j \in L : i \neq j\}}$$

It is likely that this matrix represents a fully connected graph. Because information does not always follow geodesic paths, and because we want to weight information paths, over this graph we estimate *random-walk weighted betweenness centrality* [17]. To confirm the presence of centralization, we consider the estimated centrality as the *observed centrality in location interactions*, which is compared to the *expected centrality in location interactions*. The expected centrality is the random walk betweenness centrality estimated in the interaction graph, considering edge weights as the normalized geometric mean of location populations. A considerable deviation from the expectations is a strong signal of centralization.

Document Location Representation. Our initial assumption is that each location will have several local words and hashtags that characterize it. These hashtags, among other words like place names, people names and vernacular words, will have more weight in their corresponding documents than global, non local words. Hence, we consider that a tweet talks about a particular location if its content resembles or is similar enough to the aggregated content of that location. To build a *location corpus* of $|L|$ *location documents*, we consider the set of geolocated users U_L. Each document is the aggregation of tweets originating from those locations, leaving out *replies*, *mentions* and *retweets* to avoid repeated content between different documents. We represent each location document d as a vector $\vec{d} = [w_0, w_1, \ldots, w_n]$, where w_i represents the vocabulary word i weighted according to its locality by using TF-IDF [2]:

$$w_i = freq(w_i, d) \times log \frac{|L|}{|l \in L : w_i \in l|}$$

Classifying Tweets using Location Similarity. To predict a location for a given document $\vec{d}$, we build a feature vector $\vec{f_d}$ containing the similarity of $\vec{d}$ with each location document from the location corpus. In this way, we consolidate all similarities in a single vector: $\vec{f_d} = [f_0, f_1, \ldots, f_{|L|}]$, where f_i is the cosine similarity between the document $\vec{d}$ and the location document $\vec{l_i}$:

$$cosine\_similarity(\vec{d}, \vec{l_i}) = \frac{\vec{d} \cdot \vec{l_i}}{\| \vec{d} \| \| \vec{l_i} \|}$$

We use the feature vectors and their corresponding author geolocations to train classifiers based on *Support Vector Machines* (SVM) [6] and *Naive Bayes*. We define that a prediction is correct if the location predicted for a tweet matches the author location. Although this approach may give *false positives* (a user tweets about other locations) or *false negatives* (a user tweets about the event from a generic point of view), this assumption is also made in previous work [5, 10] because the usage of the self-reported location in geolocation allows to assign only one location to every user, which we find acceptable when considering events where users are expected to have a single location.

Evaluating Geographical Diversity. To find how different classifiers behave when considering geographical diversity, we define *geographical diversity* as the normalized *Shannon entropy* [11] with respect to locations:

$$geodiversity = \frac{-\sum_{i=1}^{|L|} p_i \log p_i}{log|L|}$$

where p_i is the fraction of predictions for location i, and L is the set of locations ($|L| > 1$). To balance classifier accuracy

and geographical diversity, based on the *F-measure* [2], we define a *D-measure* as the harmonic mean between accuracy and *geographical diversity*:

$$D_\beta = (1+\beta^2) \cdot \frac{geodiversity \cdot accuracy}{(\beta^2 * geodiversity) + accuracy}$$

where β establishes the weight given to diversity: D_1 gives equal weight to accuracy and diversity, $D_{0.5}$ gives more weight to accuracy, and D_2 gives more weight to diversity.

Filtering Information Streams. Given an event of interest E and a set of related tweets T_E, we generate a summary tweet set or timeline T_θ, where T_θ contains s tweets. To maximize geographical diversity of T_θ, we consider a greedy algorithm from prior work [7]. This algorithm generates T_θ from an *information entropy* perspective, where entropy is estimated in terms of several features extracted from tweets. Since the complexity of those dimensions can be greater than those of geography (for instance, consider the number of hashtags in an event and the number of locations), the entropy contribution of these dimensions is higher than the entropy contribution of geography. Thus, diversity can still be optimal even in the absence geographical diversity.

To inject geographical diversity, we extend [7] by adding a sidelining step [16] when considering tweets for inclusion in T_θ, that is, tweets from a location previously selected in the previous iterations of the algorithms will not be considered for a given number of turns. Additionally, for user interests to be represented, we introduce popularity into the input features, which has been established as a valuable feature for tweet recommendation [3, 4].

For each tweet t, we consider a vector representation $\vec{v_t}$ with the following features:

1. *Presence of links*: whether the tweet contains an URL.
2. *Time bucket*: n^{th} time-window of 5 minutes since the start of E.
3. *Annotated hashtags*: topical information for each tweet.
4. *Geography*: defined as the location the tweet content is most likely to be about using our text classifiers.
5. Author's number of *followers*.
6. Author's *hub dimension* ($\frac{followers}{friends}$).
7. Author's *global tweet count*.
8. *Popularity*: number of times the tweet has been retweeted.

For a given tweet set T', with k different vector representations of its elements ($k \leq |T'|$), we estimate its *Shannon entropy* [11]: $H_{T'} = -\sum_{i=1}^{k} p(\vec{v_{t_i}}) \log p(\vec{v_{t_i}})$. Then, to create a summary tweet set T_θ of size s, where *turns* is the number of times a location is not considered for addition (*"it is sidelined"*) and t_l is the predicted location for tweet t, we define the following information filtering algorithm:

1. Define a dictionary mapping *sidelined* where $sidelined[l] = 0, \forall l : l \in L$.
2. Start by selecting a tweet t from the most popular ones from T_E as initial seed in T_θ. Set $sidelined[t_l] = turns$.
3. For all tweets in T_E which are not in T_θ, build a tweet set T_c, where every tweet t in T_c satisfies the following: its addition to T_θ maximizes H_{T_θ}.
4. For all tweets in T_c, leave out those where $sidelined[t_l] > 0$, and select randomly a tweet t' from the remaining most popular tweets.
5. Set $sidelined[l] = sidelined[l] - 1, \forall l : l \in L, l \neq t'_l$
6. Set $sidelined[t'_l] = turns$
7. Repeat step 3 unless $|T_\theta| = s$.

Data	#	Fraction
User Accounts	199951	–
w/ Location Text	131625	65.83%
All Tweets	886813	–
RTs	306377	34.55%
w/Mentions or Replies	252525	28.48%
w/Hashtags	274592	30.96%
w/Geo. Coordinates	66196	7.46%
Vocabulary Size	65516	–

Table 1: Our information space: main types of data crawled during the `#municipales2012` event. *w/* means *with*.

Level	# Users	# Tweets
Country	18902 (9.45%)	112321 (12.67%)
Region	420 (0.21%)	3385 (0.38%)
Province	2048 (1.02%)	15186 (1.71%)
Municipality	53436 (26.72%)	355224 (40.06%)
Geolocated	74806 (37.41%)	486116 (54.82%)

Table 2: Number and geographical level of geolocated accounts using the self-reported location.

3. CASE STUDY: CHILE

The administrative locations of Chile are defined according to the following hierarchy: *municipality* → *province* → *region* → *country*. In our work, we consider the 15 Chilean regions, as is the level at which Chilean centralization is characterized [23]. The most populated and centralized region is *Región Metropolitana* (*RM*).

Dataset. Our dataset is composed of tweets crawled on October 28, 2012, related to the municipal elections held in Chile. The event had a distinctive hashtag, `#municipales2012`, which among other related hashtags (e.g. `#túdecides`), keywords (e.g. `vote`), location and candidate names were used as queries to the *Twitter Streaming API*.[1] Table 1 gives an overview of the dataset after removing unrelated tweets and tweets not in Spanish. The fraction of tweets with geographical coordinates is very low (7.46%) and the fraction of tweets with hashtags is less than a third (30.96%).

Virtual Population. To initialize the gazetteer, we loaded a list of 1978 toponyms from previous work [9]. Table 2 contains the number of users and tweets per location level. Only 37.41% of the participating accounts could be geolocated, although those accounts produce 54.82% of the event content, a reasonable amount to build the corpus of location documents. Figure 1 (left top) shows the number of accounts per region, showcasing the imbalance in population distribution. The mean rate of regional twitter accounts per 1000 inhabitants is 2.65, indicating that the proportion of accounts in each region is similar (see Figure 1, top right). To explore the representativity of the sample, we estimate the *Pearson product-moment correlation coefficient* between virtual and physical population, and between account rate and household Internet Access Rate in Chile [14]. Because of population imbalance, we estimate the correlation of the logarithms of population, obtaining 0.95 ($p < 0.01$, Figure 1 bottom left). The user and household Internet access rates correlate at 0.68 ($p < 0.01$, Figure 1 bottom right).

[1] `https://dev.twitter.com/docs/streaming-apis`, accessed 2-July-2014.

Region	Tweet Share	Incoming Share	Import %	RM Import %	Export %	RM Export %	Exp./Imp.	Expected Centrality	Observed Centrality
I	2.52	1.90	16.01	9.14	36.81	29.64	3.06	0.04	0.01
II	2.71	1.60	26.66	13.43	56.72	46.66	3.60	0.05	0.02
III	0.72	0.44	45.11	24.58	66.47	54.66	2.41	0.04	0.01
IV	2.34	2.13	54.04	35.90	58.04	48.36	1.18	0.06	0.02
V	10.06	5.77	46.02	30.29	69.07	60.64	2.62	0.10	0.09
VI	1.43	0.67	64.51	38.96	83.44	71.52	2.77	0.07	0.01
VII	2.78	1.39	38.62	24.04	69.25	57.97	3.58	0.07	0.02
VIII	8.23	5.25	37.63	23.80	60.22	51.51	2.51	0.10	0.07
IX	3.19	1.89	33.83	18.92	60.75	49.36	3.03	0.07	0.03
X	2.39	1.28	35.95	20.31	65.60	53.37	3.40	0.07	0.02
XI	0.33	0.14	61.76	30.15	83.33	68.59	3.10	0.02	0.00
XII	1.13	1.03	56.81	36.48	60.28	50.65	1.15	0.03	0.01
XIV	2.46	1.95	32.54	20.11	46.71	37.92	1.82	0.04	0.03
XV	0.85	0.59	50.80	27.45	65.67	53.48	1.85	0.03	0.01
RM	**58.87**	**73.97**	**29.12**	**70.88**	**10.95**	**89.05**	**0.30**	**0.19**	**0.76**

Table 3: Information trade in terms of proportion of outgoing and incoming tweets for the Chilean regions. Centrality is estimated from the adjacency matrix of interactions.

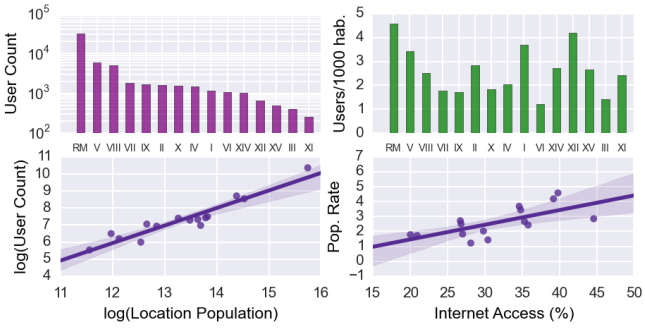

Figure 1: Top: Distributions of population according to Chilean regions (left) and user rates per 1000 inhabitants (right). Bottom: linear regressions of: logarithms of physical population with Twitter accounts (left), Internet access rate with Twitter account rate (right).

Therefore, we consider our sample spatially representative of the physical population at the regional level.

Location Interactions. As defined in Section 2, we build the adjacency matrix of 1-way interactions and its corresponding graph. We observe in Table 3 that *RM* produces the majority of tweets (58.87%) and receives most of the incoming interactions (73.97%). The ratio of exported/imported tweets shows an interesting pattern: the only region with a ratio below 1 is *RM*, suggesting the presence of centralization. Then, we estimate *random walk betweenness centrality* [17] considering the edge weights defined in Section 2. Table 3 shows both expected and observed centralities for all locations: *RM* is, indeed, the most central location (0.76 expected, 0.19 observed), having the only increased and highest difference between observed and expected centralities. Therefore, in the context of our dataset, there is centralization in location interactions.

Content Location Classifiers. After building the location corpus, we apply TF-IDF to find the most discriminating hashtags and keywords for each location. To remove noise, when building location documents we discarded hashtags and keyword appearing in less than 5 different tweets. Most of the found discriminative keywords can be categorized in: *a)* toponyms, like `#laserena` (IV) and `coyhaique` (XI); *b)* names of candidates from the corresponding region: `#soria` (I) and `arellano` (VI); and *c)* local adaptations of event hashtags: `#municipalesmag` (XII) and `#municipalesfm` (V). This result validates our assumptions about local vocabulary.

Using the location documents and the set of tweets from the geolocated users, we build the feature vectors corresponding to each tweet. We evaluated the following classifiers using a 10-fold stratified cross-validation: SVM Linear Kernel (*one versus one* multiclass strategy), SVM Linear Kernel (*one versus all* multiclass strategy) [21], SVM RBF Kernel and Naive Bayes. We divided the set of tweets from geolocated users in 10 groups, maintaining the proportions of users' tweets in each group, and then ran 10 iterations to evaluate the classifiers. In each iteration we trained each classifier using 9 tweet groups and tested predictions with the remaining group. In this way, each tweet was used nine times for training and one time for evaluation. We did not consider retweets and replies to avoid duplicate tweets in training and evaluation sets. Then, we estimated the geographical diversity of the set of predictions of each classifier, and calculated the *D-measure* at $\beta = \{0.5, 1, 2\}$. To evaluate results of our approach, we considered the following baselines: *a) Trivial Classifier*, which predicts the most common location in the dataset (*RM*); *b) Best Cosine Similarity*, which predicts the location with the highest cosine similarity between the tweet content and the location documents, as in [9]; and *c) SVM* and *Naive Bayes* classifiers trained with *bag of words*, with vocabulary size 65516. Results are reported in Table 4.

In terms of accuracy, SVM has the best performance in both scenarios (using similarity features and bag of words), which aligns with previous work on imbalanced populations [22]. However, not all classifiers show diversity: some of them have entropy 0, which means that they are behaving in the same way as the trivial classifier. Although geographical diversity is important in our context, accuracy is needed to avoid inconsistent results. Thus, we consider $D_{0.5}$, in which the best scores are for *Cosine Similarity* (0.60) and *SVM Linear1vsA* (0.58). We believe *SVM Linear1vsA* is a better solution, as *Cosine Similarity* has worse performance when considering finer location granularity [9] and SVM has proven to be robust at different levels [22]. Note that even without our features *SVM Linear1vsA* had a considerable amount of diversity (but still lesser than with our approach).

Approach	Acc.	Geo. Div.	D_1	$D_{0.5}$	D_2
SVM Linear1vs1	0.68	0.34	0.45	0.56	0.38
SVM RBF	0.68	0.34	0.45	0.56	0.38
SVM Linear1vsA	0.68	0.38	0.49	0.58	0.41
Naive Bayes	0.58	0.00	–	–	–
Cosine Similarity	0.62	0.54	0.58	0.60	0.56
W-SVM Linear1vs1	0.58	0.00	–	–	–
W-SVM RBF	0.58	0.00	–	–	–
W-SVM Linear1vsA	0.68	0.27	0.39	0.52	0.31
W-Naive Bayes	0.58	0.00	–	–	–
Trivial	0.58	0.00	–	–	–

Table 4: Evaluation results at regional level of our classifiers using a 10-fold stratified cross validation. Classifiers prefixed with *W-* use normalized word counts, while other classifiers use TF-IDF weighting according to locations.

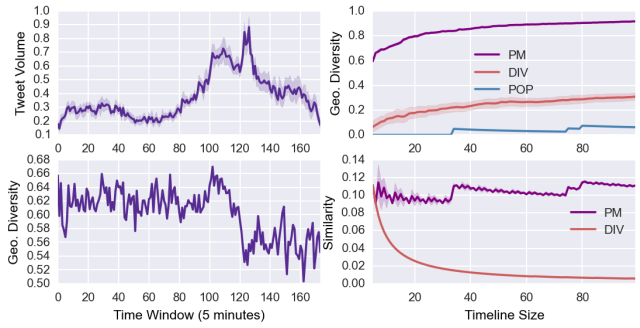

Figure 2: Tweet volume during the event (top left) and corresponding geographical diversity (bottom left). Normalized geographical diversity for timeline sizes between 5 and 100 (top right). Jaccard similarity between generated timelines using filtering and the popularity sampling for timeline sizes between 5 and 100 (bottom right).

Temporal Geographical Diversity. Figure 2 shows tweet volume (top left) and geographical diversity (bottom left) of the dataset: from morning to afternoon, activity increased steadily as the elections were being held. At night, specific events regarding unexpected results raised the level of activity above expectation. Before this peak, geographical diversity had a mean value of 0.62; after, the mean geographical diversity was 0.55, a decay explained by the unexpected defeat of several candidates in some locations, shifting the discussion in a natural way towards fewer locations.

Filtering Algorithm. Because geographically diverse content exists, we evaluate if our information filtering algorithm produces geographically diverse summary timelines. We consider the following baselines against our *Proposed Method* (`PM`): *B1) Popularity Sampling* (`POP`): we select the s most popular tweets in terms of retweets; *B2) Diversity Filtering* (`DIV`): an implementation of [7] considering the tweet predicted location as a geographical feature. We estimate the $s = 100$ most popular tweets for `POP`, and run `DIV` and `PM` a hundred times (`POP` runs only once because the outcome is always the same for the same input). At every timeline size $i \in [5, 100]$ we estimate: *1)* the geographical diversity of `POP`, `DIV` and `PM`; *2)* the Jaccard similarity between `DIV` and `POP`, and between `PM` and `POP` (defined as $J(A,B) = \frac{|A \cup B|}{|A \cap B|}$). Our algorithm consistently shows greater geographical diversity than `POP` and `DIV` (Figure 2 top right), indicating that our sidelining step produces the needed effect of *geodiversification*. The lack of geographical diversity in POP could be a consequence of centralization, as even on the imbalanced population distribution at least 41.13% of the tweet share is generated outside of RM (see Table 3). In terms of the similarity with POP, our method is much more similar than DIV (Figure 2 bottom right). Because we modified the baseline algorithm from [7] to start with one the most popular tweets instead of a random selection, the initial similarity is the same as in our method. However, as the timeline size increases, the similarity to POP tends to become 0, whereas our method maintains a similarity between 0.09 and 0.12 (Figure 2 bottom right). Hence, our method has a stronger representation of user interests than DIV.

4. DISCUSSION

System biases affecting how people behave should be considered in system design. In this line, we explored how *systematic biases from the physical world are reflected on microblogging platforms*. In the case of Chile, we confirmed that centralization happens in Twitter when the population from the physical world is centralized (Table 3).

Then, we proposed text features for geolocation and found that *similarity features help classification in biased scenarios* (see Table 4). In this aspect, when using SVM the *one vs all* approach [21] is better than *one vs one*, as it has greater diversity with our features and it does not become overfitted when using the *bag of words* approach. In terms of scalability, similarity features are orders of magnitude smaller than the vocabulary size in the *bag of words* approaches. Moreover, a classifier trained over similarities should tolerate new vocabulary without needing re-training, as it is built over how similar a given text is to the location documents instead of word distributions.

Finally, to balance event-summary timelines from a geographical perspective we defined an information filtering algorithm based on previous work. In an off-line evaluation, our algorithm outperformed the baselines in terms of geographical diversity while still maintaining desirable properties such as information diversity and representation of user interests.

Limitations. Critics might rightly say that other approaches could have led to better precision and higher recall than our ad-hoc gazetteer. We analyzed the representativity of the sample of geolocated users and found that it was spatially representative of the population. The focus of our work is not o identify the exact user geolocation, but to use the latter to promote diverse timelines. Our approach is thus sufficient for this work.

Future Work. It remains to be seen if centralization is stronger at other geographical levels, as well as the behavior of the similarity features when classifying tweets. Because our results need to be explained from user perspectives in qualitative terms, a long-term longitudinal study is needed to evaluate the real effect of incorporating these results (e.g. promoting geographical diversity) into microblogging platforms and their user interfaces.

Acknowledgments. We thank Bárbara Poblete for many fruitful discussions about this work. This work was partially funded by Grant TIN2012-38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain.

5. REFERENCES

[1] Ricardo Baeza-Yates, Christian Middleton, and Carlos Castillo. The Geographical Life of Search. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 252–259. IET, 2009.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern information retrieval: the concepts and technology behind search, 2nd. Edition*. Addison-Wesley, Pearson, 2011.

[3] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.

[4] Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670. ACM, 2012.

[5] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 759–768. ACM, 2010.

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[7] Munmun De Choudhury, Scott Counts, and Mary Czerwinski. Identifying relevant social media content: leveraging information diversity and user cognition. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 161–170. ACM, 2011.

[8] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. Cultural Dimensions in Twitter: Time, Individualism and Power. In *International AAAI Conference on Weblogs and Social Media*, 2013.

[9] Eduardo Graells-Garrido and Bárbara Poblete. #Santiago is not #Chile, or is it?: a model to normalize social media impact. In *Proceedings of the 2013 Chilean Conference on Human-Computer Interaction*, pages 110–115. ACM, 2013.

[10] Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 237–246. ACM, 2011.

[11] Lou Jost. Entropy and diversity. *Oikos*, 113(2):363–375, 2006.

[12] Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P Gummadi. Geographic dissection of the Twitter network. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[13] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101. ACM, 2006.

[14] Gobierno de Chile Ministerio de Desarrollo Social. CASEN Survey. `http://observatorio.ministeriodesarrollosocial.gob.cl/casen_obj.php`, 2011. [In spanish; Online; accessed 2-July-2014].

[15] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11), Barcelona, Spain*, 2011.

[16] Sean A Munson, Daniel Xiaodan Zhou, and Paul Resnick. Sidelines: An Algorithm for Increasing Diversity in News and Opinion Aggregators. In *Third International AAAI Conference on Weblogs and Social Media*, 2009.

[17] Mark EJ Newman. A measure of betweenness centrality based on random walks. *Social networks*, 27(1):39–54, 2005.

[18] Bárbara Poblete, Ruth García, Marcelo Mendoza, and Alejandro Jaimes. Do all birds tweet the same?: characterizing Twitter around the world. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1025–1030. ACM, 2011.

[19] Daniele Quercia, Licia Capra, and Jon Crowcroft. The social world of Twitter: Topics, geography, and emotions. In *The 6th international AAAI Conference on weblogs and social media*, 2012.

[20] Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing Microblogs with Topic Models. In *ICWSM*, 2010.

[21] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.

[22] Dominic Rout, Kalina Bontcheva, Daniel Preoţiuc-Pietro, and Trevor Cohn. Where's@ wally?: a classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pages 11–20. ACM, 2013.

[23] Wikipedia. Centralismo en Chile — Wikipedia, The Free Encyclopedia. `http://es.wikipedia.org/wiki/Centralismo_en_Chile`, 2013. [In spanish; Online; accessed 2-July-2014. Title translation: "Centralization in Chile"].

[24] Wikipedia. Elecciones municipales de Chile de 2012 — Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/wiki/Chilean_municipal_election,_2012`, 2013. [Online; accessed 2-July-2014].

[25] Sarita Yardi and Danah Boyd. Tweeting from the Town Square: Measuring Geographic Local Networks. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.