User-driven geolocated event detection in social media – Supplementary material

Anes Bendimerad, Marc Plantevit, Céline Robardet, Sihem Amer-Yahia

1 User-driven discovery based on simulated users

In this section, we provide supplementary experiments of the user-driven approach. We simulated virtual users who, depending on their "own interest", tend to prefer events of a given type. We evaluate their "satisfaction" according to the number of events that the system presents to them that they are inclined to like. Thus, virtual users having topics and/or location preferences are simulated and the experiment consists in studying the number of liked events when considering or not user feedback during the event discovery process.

To that end, we first extracted events in the three real datasets[1] and retain those spanning at least over 2 timestamps (the set $\mathcal{E}$). Then, we manually annotated the events. The tags used for the annotation are Topics $=$ $\{$Business/Economics, Politics, Science/Technology, Art/Culture, Celebration, Music, Sport, Accident/Disaster$\}$. Each event P can be annotated with several topics and the function $\text{Tag}(P,\tau) \rightarrow [0,1]$ expresses the importance of the topic for the event (with $\sum_{\tau \in \text{Topics}} \text{Tag}(P,\tau) = 1$). Some detected events did not match to any category and the obtained distributions are presented in Table 1.

dataset	#	Art/Culture	Music	Celebration	Sport	Politics	Business
NYC	800	97	89	212	87	88	1
LA	489	157	70	18	30	0	7
London	353	120	32	8	53	3	36

TABLE 1
Distribution of events according to the topics.

A virtual user u prefers a specific topic, or location, or both of them. The function $\mathcal{P}_u(P)$ captures the preferences of the user u for the event P. It is defined according to 3 cases:

- If τ is the preferred topic of u, $\mathcal{P}_u(P) = \text{Tag}(P,\tau)$
- If ℓ is the preferred location of u, we consider that u is interested in events at a distance from ℓ at most equal to L and $\mathcal{P}_u(P) = \max(\frac{L - \text{dist}(\ell,P)}{L}, 0)$. Based on the surface area of the cities, we choose $L = 5km$ for New York, and $L = 10km$ for Los Angeles and London.
- If u has both topic and location interests, $\mathcal{P}_u(P) = \frac{\text{Tag}(P,\tau)+\max(\frac{L-\text{dist}(\ell,P)}{L},0)}{2}$

Topics with fewer than 20 events were discarded. For the preferred locations, we consider several well-known places for each city[2]. Finally, to be able to automatically annotating computed events on these datasets, we used the Tag function to annotate hashtags and locations (for $x = h, v$, $\text{Tag}'(x,\tau) = \sum_{P \in \mathcal{E} \ s.\ t.\ x \in P} \text{Tag}(P,\tau)$) and then use them to automatically annotate events ($\text{Tag}^{\star}(P,\tau) = \sum_{x \in P} \text{Tag}'(x,\tau)$).

To evaluate how `SIGLER-Cov` and `SIGLER-Samp` effectively discover user-driven events, we simulate the same interactive process that we did with the real users in Section 6.4, but with virtual users this time.

Fig. 1 presents results of these experiments when simulating between 19 and 29 virtual users depending on the number of considered locations and topics on each dataset. For each dataset, we show boxplots of the average number of likes using `SIGLER-Cov` and `SIGLER-Samp` in the data and user-driven settings. We can observe that (1) the average number of likes in the user-driven setting is always greater than the one in data-driven configuration. This difference is considered as significant by the Wilcoxon and the Nemeny post-hoc [Dem06] tests (the later is shown on Fig. 1.(2) results obtained by `SIGLER-Samp` are below those obtained by `SIGLER-Cov`, and the difference is significant according

- *A. Bendimerad and C. Robardet are with University of Lyon, INSA Lyon, CNRS UMR 5205.*
- *M. Plantevit is with University of Lyon, University Lyon 1, CNRS UMR 5205*
- *S. Amer-Yahia is with University of Grenoble Alpes, CNRS.*

[1] To obtain around 800 events on NYC, we used $\delta = 40$, and to obtain around 400 events on LA and London, we used $\delta = 15$. On all datasets we set $\texttt{minCov} = 0.8$.

[2] **NYC**: Barclays Center, Javits Center, Madison Square Garden and Metlife Stadium; **LA**: City Hall, DisneyLand, Museum of Art and Rose Bowl Stadium; **London**: City of London, Royal Albert Hall, Soho Theatre and Wembley Stadium.

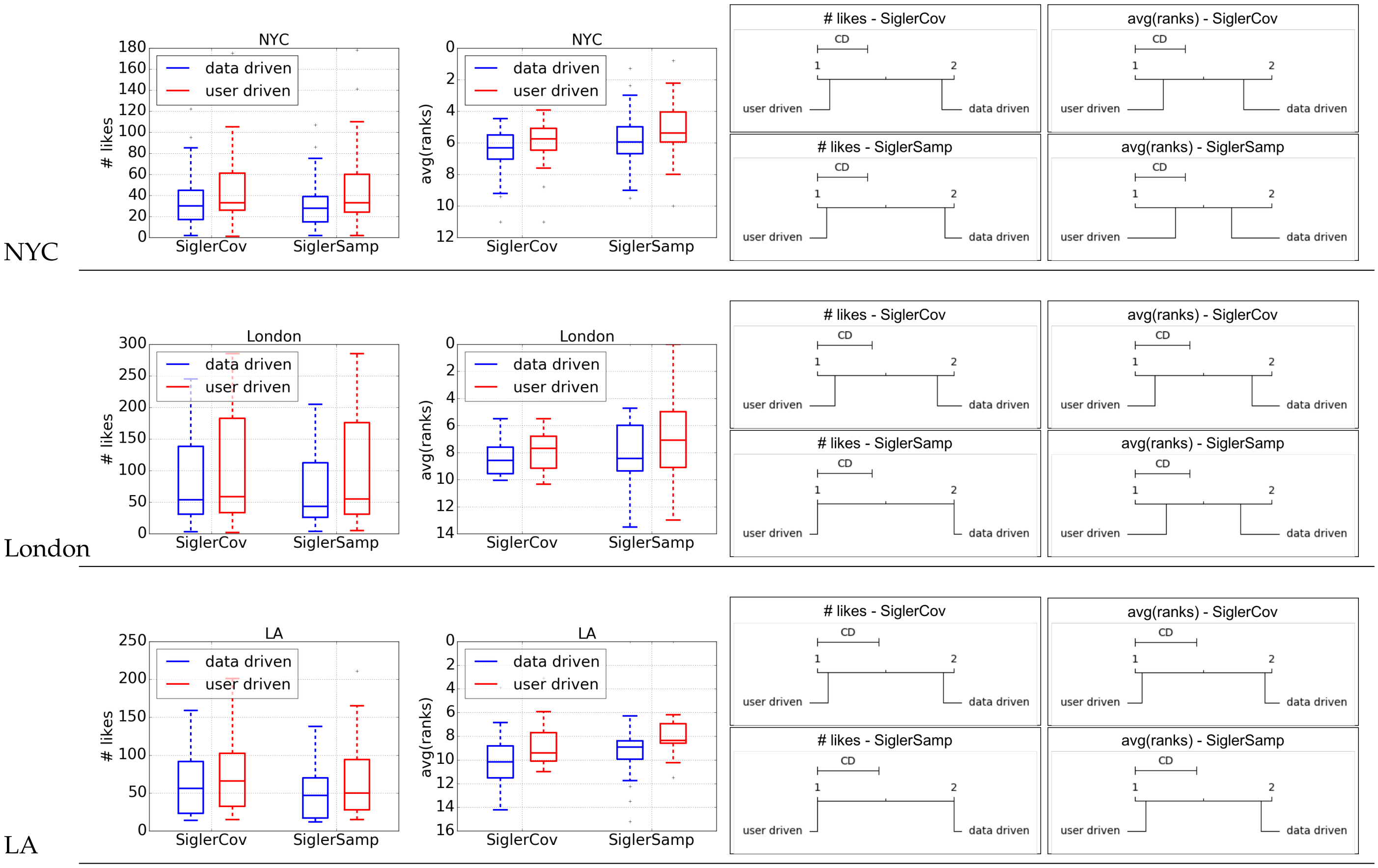

Fig. 1. Virtual user-driven geo-spatial event detection with SIGLER-Cov and SIGLER-Samp. Number of likes (1st column) and average ranks of events liked by both data and user-driven settings (2nd column), Nemenyi tests on number of likes and average ranks (3rd and 4th columns).

to the Wilcoxon test, except for the NYC dataset, where the number of likes is considered to be similar. Indeed, SIGLER-Cov is more exhaustive and finds more events than SIGLER-Samp, which explains this point.

Notice that in this simulation between 300 and 800 events are presented to the virtual-users who, on the contrary to human users, have the ability to evaluate all of them. As already mentioned above, the order in which the events are presented to the users is essential. In order to evaluate this point, Fig. 1 presents the average ranks of events liked by both data and user-driven settings. Clearly the setting for which the liked events are ranked first is advantageous in real situations. We can observe that the average rank is always lower in the user-driven setting than in the data-driven one. This difference is considered significant by the Wilcoxon test on all datasets. This is also confirmed, in a more visual way, by the Nemenyi tests [Dem06] displayed on the figure. When comparing the average ranks obtained by SIGLER-Cov and SIGLER-Samp, it appears that the later obtains significantly better ranks (smaller) for liked events that the former. Thus SIGLER-Samp identifies fewer events, but that are of high quality for users.

This simulation with virtual users allows us to conclude that the user-driven setting makes it possible to identify more relevant events than the data-driven one, whether in terms of quantity and quality. Besides, SIGLER-Samp identifies fewer geolocated events than SIGLER-Cov, but they are of better quality.

2 Comparative study in real world dataset

Based on synthetic data, we have studied in Section 6.2 the ability of our approach to detect local events, and we compared it with other state-of-the-art methods. Ideally, one would prefer to use real world datasets to perform such evaluation. However, we do not have the ground truth of the studied real world data. This makes it very hard to achieve an objective comparison on them. Nevertheless, we show in Table 2 the top 10 events returned by SIGLER-Cov, MED, and GeoBurst, on the first 10k tweets of NYC dataset, and we make a discussion about them. The number of tweets is limited to only 10k, in order to be able to compare with MED which has scalability issues.

We can notice that there are some similar results of SIGLER-Cov with those of MED and GeoBurst. In fact, SIGLER-Cov and MED have both returned the New York Comic Con[3] (1, 4 and 8 in SIGLER-Cov, 3 and 4 in MED), Beyoncé Concert[4] (2 in SIGLER-Cov, 5 and 8 in MED), and Taylor Mac concert[5] (5 in SIGLER-Cov and 10 in MED) .

[3] https://goo.gl/BR7kgp
[4] https://goo.gl/FrZEBu
[5] https://goo.gl/9pM6z3

#	SiglerCov		MED		GeoBurst	
	time	top 5 terms	time	top 5 terms	time	top 5 terms
1	0h to 23h	**nycc, nycc2016, cosplay, ny_comic_con, newyorkcomiccon.**	0h to 23h	nyc, newyork, love, manhattan, saturday.	15h to 23h	**nilerodgers, foldfest, bettemidler, foresthillsstadium, walkeratconcert.**
2	0h to 21h	**formationworldtour, beyonce, formationtour, beyhive, theformationworldtour.**	0h to 23h	newyork, job, hiring, newyorkcity, photo.	15h to 21h	**raniahatoum, thelondonnyc, tripleb, blackbridalbliss, bridalgown.**
3	6h to 18h	**rnrbrooklyn, runrocknroll, halfmarathon, brooklynwerunhard.**	0h to 23h	**nycc2016, cosplay, nycc, newyorkcomiccon, wonderwoman.**	18h to 23h	**intercoiffure, wella, icamoments, nerolisalonspa, wellapro.**
4	12h to 21h	**smashingnycc, nigelthornberry, nigel, smashing, wildthornberries.**	0h to 23h	**nycc, ny_comic_con, cosplay, comiccon, marvel.**	9h to 15h	ridetheferry.
5	12h to 21h	**24decadehistoryofpopularmusic, sawtaylormac, 24decades, marskado, afraidoffun.**	0h to 23h	**formationworldtour, beyonce, beyhive, metlifestadium, beyonce.**	18h to 23h	**sturgillsimpson, kingsbklyn, asailorsguidetoearth.**
6	15h to 23h	**greenday, websterhall, revrad, saturdaynight, 90s.**	0h to 23h	**brooklyn, bushwick, williamsburg, sigurros, music.**	0h to 12h	**elitefridays, cityscapesny, imsobx, reposting, cityscapesnyc.**
7	18h to 21h	**dosgualas, livvinyl, monies, freeze, megaman.**	0h to 23h	repost, montanoy27, regram, alofokemusicnet, parkslopemoms_.	18h to 23h	**descendents.**
8	12h to 18h	**ronswwadventure, 75thanniversary.**	0h to 23h	**formationtour, beyonce, theformationworldtour, nj, kendricklamar.**	0h to 15h	**50cent, dozadrumdealer, mynameisjuan, industrykilla, narcotechs.**
9	15h to 21h	**foresthillsstadium, nilerodgers, fold, bettemidler, foresthills.**	3h to 23h	foodporn, food, foodie, yummy, eeeeeats.	0h to 12h	**deadrabbitnyc.**
10	18h to 21h	**knicks, nets, nyknicks, brooklynnets, preseason.**	0h to 23h	**24decadehistoryofpopularmusic, sawtaylormac, 24decades, proofoflifenumber, marskado.**	0h to 23h	**doomocracy, pedroreyes, doomacracy, creativetime.**

TABLE 2

Top 10 events returned by SIGLER-Cov, MED, and GeoBurst, in NYC dataset, for the first 10k tweets (the day of 8 Oct. 2016). True positive events are market in bold while false negative events are not.

Both SIGLER-Cov and GeoBurst identified the FOLD Festival of Nile Rodgers[6] (9 in SIGLER-Cov and 1 in GeoBurst). However, the rest of top results of GeoBurst are different from the ones of other approaches. GeoBurst seems to give more importance to small events. In fact, each of the top 10 events of GeoBurst contain at most 10 tweets, while the number of tweets in top results of SIGLER-Cov (resp. MED) varies between 43 and 3k (resp. between 42 and 1280).

We believe that the results 1, 2, 7, and 9 of MED, and the result 4 of GeoBurst are not relevant. Indeed, they are defined with terms that do not correspond to any real life event (e.g., nyc, job, hiring, foodporn, etc.). Concretely, The terms ”nyc, newyork, love, manhattan, job, hiring, newyorkcity, photo, repost” are very frequent in New York dataset. Each of them appear at least 30 times in 90% of the days. The term ”saturday” is frequently used in Saturday (more than 30 times in 80% of cases). The terms ”food, foodie, yummy, eeeeeats” also appear in a large number of posts where people want to share their feeling about some food experience. Each of them is used at least 7 times in 50% of the days. The term ”ridetheferry” is used by people who pass by the NY Waterway Ferry. It occurs between 1 and 6 times in 22 different days.

References

[Dem06] Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[6] https://goo.gl/1htnhk