Hashtag Similarity Based on Laplacian Eigenvalue Spectrum

Bartłomiej Starosta [0000−0002−5554−4596]

Mieczysław A. Kłopotek [0000−0003−4685−7045]

Sławomir T. Wierzchoń[0000−0001−8860−392X]

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
barstar,klopotek,stw@ipipan.waw.pl

Abstract. *Hashtags play nowadays an important role in the current social media world. They are usually deemed to represent topics of e.g. tweets. As the number of hashtags is growing, an overview of the information flow requires some method of grouping these hashtags. The grouping requires a similarity measure. In this paper we propose a novel measure of similarity between hashtags based on the Graph Spectral Analysis.*
Keywords: *Graph Spectral Analysis, combinatorial Graph Laplacian, eigenvalue spectrogram based similarity, artificial intelligence*

1. Introduction

The so called Graph Spectral Analysis (GSA) represents a novel way of looking into relationships between data objects that are characterized by mutual similarity measures, and hence can be best described by a graph with weights equal to these similarities. The similarity matrix is transformed to e.g. combinatorial Laplacian, which in turn is subject to eigendecomposition. Eigenvectors constitute a new coordinate system into which the data objects are embedded and thus may be subject of distance-based data clustering or data classification methods [1, 2], also with hashtags [3]. The main stream of research concentrates on usage of a carefully selected subset of eigenvalues and corresponding eigenvectors.

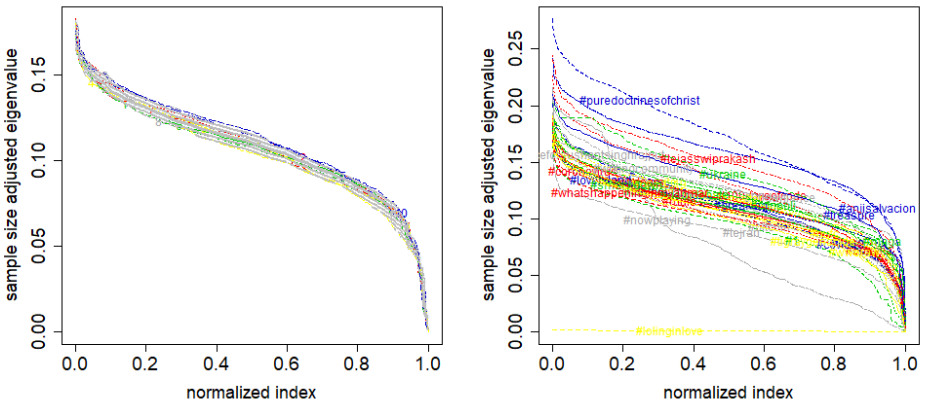

Figure 1. Normalized spectrograms for samples of (left:) one single hashtag, (right:) various hashtags

Our experiments (in Sect.3) have shown, however, that there exists a possibility to use the entire eigenvalue spectrogram as a way to characterize classes within the aforementioned weighted graph of objects and consequently, the similarity of the spectrograms is usually related to the similarity of the classes themselves. The paper explains our methodology in brief in Sect.2. It is based on the observation that spectra of combinatorial Laplacian of random subsamples of the same class can be down-scaled to overlap, while those from different classes do not.

2. The Method

Let S be a similarity matrix between pairs of items (e.g. tweets). It induces a graph whose nodes correspond to the items. A(n unnormalised) or combinatorial Laplacian L corresponding to this matrix is defined as

$$L = D - S\,, \tag{1}$$

where D is the diagonal matrix with $d_{jj} = \sum_{k=1}^{n} s_{jk}$ for each $j \in [n]$.

Let its eigenvalues be in non-decreasing order $0 = \lambda_1 \leq \cdots \leq \lambda_n$.

We proposed a function $\lambda_{CLSSAL} : [0,1] \to \mathbb{R}$ in such a way that

$$\lambda_{CLSSAL}\Big(\frac{n-i}{n-1}\Big) = \frac{\lambda_i}{n} \ . \tag{2}$$

The linear interpolation is applied in-between.

Based on the above assumption, we can compute a "distance" between a given new sample and the elements of a class as the area between the λ_{CLSSAL} curves. So if the first subgraph $G1$ is characterized by $\lambda_{CLSSAL,G1}$ curve, and the second subgraph $G2$ is characterized by $\lambda_{CLSSAL,G2}$ curve, then the dissimilarity is computed as

$$dissim(\lambda_{CLSSAL,G2}, \lambda_{CLSSAL,G1}) = \int_0^1 |\lambda_{CLSSAL,G2}(x) - \lambda_{CLSSAL,G1}(x)| dx \tag{3}$$

3. Experiments

We investigated this phenomenon for a small collection of hashtags extracted from Twitter tweets. Their names are listed in the first column of the Table 1. We constructed a graph of tweets having only one hashtag from this list, where the weights of the tweets are computed as cosine measure in the bag-of-words vector space.

We investigated two types of subgraphs of this graph: subgraphs that include all objects of the same hashtag and subgraphs of such graphs.

For each of the subgraph we computed the combinatorial Laplacian according to equation (1). Then the function $\lambda_{CLSSAL}()$ was created for each subgraph based on the equation (2). Finally, the dissimilarity between the spectrograms was computed according to equation (3).

Fig.1, left, represents overlapped diagrams of functions $\lambda_{CLSSAL}()$ of ten samples of tweets belonging to the same hashtag. It turns out that the spectrograms of the subsets of the same hashtags are quite close to one another.

Fig.1, right, represents overlapped diagrams of functions $\lambda_{CLSSAL}()$ of 34 samples of tweets belonging to the various hashtags listed in the first column of the table 1. It turns out that the spectrograms of the subsets related to different hashtags may differ even substantially.

hashtag	s.hashtag	min.dist	avg.dist	std.dist	subsamp.dist	subsamp.err	rel.subsamp.dist
#1	#tejran	0.008004704	0.02757048	0.02010759	0.003701717	0.002201124	0.1342638
#100daysofcode	#treasure	0.006955492	0.025834407	0.02107681	0.005498188	0.002859725	0.212827
#90dayfiance	#maga	0.005161067	0.02306229	0.02111358	0.003688091	0.001489804	0.1599187
#aewdynamite	#demdebate	0.001579806	0.01779384	0.02137026	0.003271367	0.001303339	0.1838483
#anjisalvacion	#tejasswiprakash	0.01082191	0.04816793	0.02440739	0.003400781	0.00184204	0.07060261
#auspol	#coronavirus	0.001122648	0.01758666	0.02167454	0.002911198	0.002186825	0.1655344
#bbnaija	#whatsh...mar	0.004470872	0.02170779	0.02088629	0.001556021	0.0008715957	0.07168028
#bitcoin	#whatsh...mar	0.001994543	0.02048817	0.02139435	0.002381747	0.00131487	0.1162499
#blacklivesmatter	#demdebate	0.001858698	0.01791834	0.02139373	0.002099182	0.001207379	0.1171527
#breaking	#just...ajput	0.001608261	0.01806659	0.02094577	0.005108423	0.003956473	0.2827553
#cdnpoli	#covid_19	0.002422555	0.01907016	0.02120027	0.003914232	0.002502598	0.2052542
#coronavirus	#auspol	0.001122648	0.0176662	0.02168422	0.001548446	0.0005699945	0.08765021
#covid	#coronavirus	0.001857122	0.01755118	0.0215299	0.00517836	0.002254856	0.2950434
#covid_19	#cdnpoli	0.002422555	0.01823082	0.02124761	0.00580463	0.004045009	0.3183966
#covid19	#wweraw	0.001685603	0.01865511	0.02165203	0.002037235	0.000940167	0.1092052
#demdebate	#aewdynamite	0.001579806	0.01805983	0.02133996	0.002643427	0.001205658	0.1463705
#endsars	#blacklivesmatter	0.002633596	0.01757168	0.02135928	0.003110856	0.001541862	0.1770381
#justice...put	#breaking	0.001608261	0.01802846	0.0209234	0.004319812	0.00216075	0.2396107
#lolingimlove	#nowplaying	0.08095599	0.1135585	0.01869101	0.0003761939	1.161857e-05	0.003312778
#loveisland	#demdebate	0.003870494	0.01846769	0.02116395	0.001906628	0.0009974423	0.1032413
#maga	#cdnpoli	0.004050012	0.0207609	0.02121451	0.006509118	0.003850841	0.3135278
#mufc	#coronavirus	0.001592204	0.01805593	0.02168119	0.002320167	0.0006874776	0.1284989
#nowplaying	#1	0.03017543	0.04539396	0.01372244	0.01041429	0.004262152	0.2294201
#nufc	#mufc	0.001667688	0.01831197	0.02156489	0.003933591	0.0008943253	0.2148098
#puredoct...christ	#anjisalvacion	0.01333359	0.06026725	0.02505434	0.004094367	0.0016304	0.06793685
#smackdown	#auspol	0.002635276	0.01869505	0.02131051	0.004017606	0.001469838	0.2149021
#tejasswiprakash	#ukraine	0.009322752	0.03850018	0.02322012	0.004882715	0.002870357	0.1268232
#tejran	#1	0.008004704	0.03201366	0.01932727	0.003926262	0.0015748	0.1226433
#tigraygenocide	#bitcoin	0.003517509	0.02141553	0.02096667	0.004367564	0.002388737	0.2039438
#treasure	#ukraine	0.00570973	0.02789071	0.02178642	0.003076978	0.0009072463	0.1103227
#ukraine	#treasure	0.00570973	0.03094697	0.02215523	0.005446961	0.001946473	0.1760095
#whatsh...myanmar	#bitcoin	0.001994543	0.02059761	0.02131954	0.003152089	0.00101502	0.1530318
#writingcommunity	#maga	0.004232268	0.02220214	0.02099251	0.00568054	0.003211025	0.2558555
#wweraw	#covid19	0.001685603	0.01941004	0.02155251	0.005705186	0.003549209	0.2939297

Table 1. Closeness of hashtags based on eigenvalue spectrum

Table 1 shows more details of dissimilarities between the chosen tags. The column `avg.dist` presents the average dissimilarity of the given hashtag from the remaining ones, while `std.dist` shows the standard deviation of dissimilarity. The column `s.hashtag` represents the closest hashtag, with `min.dist` being the dissimilarity to it. As a contrast, `subsamp.dist` represents the average dissimilarity to 100 samples from the same hashtag, `subsamp.err` being the standard deviation of this measure. `rel.subsamp.dist` is the quotient of `subsamp.dist` / `avg.dist`.

`rel.subsamp.dist` demonstrates that in fact the samples from the same hashtag are closer to one another than to other hashtags.

The hashtag *#lolinginlove* seems to be most distant from all the other hashtags on average, while *#blacklivesmatter* seems to be close to many other hashtags from the list. The hashtag *#puredoctrinesofchrist* seems also to be distant from the other, though it is quite near to *#anjisalvacion*. *#covid* has a characteristic quite similar to *#coronavirus*.

Based on the dissimilarity matrix, most dissimilar hashtags were identified as follows: The first one was that with the highest sum of dissimilarities to other hashtags. The other were added with the highest sum of dissimilarities to those already chosen. The following list of hashtags was obtained in this way: *#lolinginlove*, *#puredoctrinesofchrist*, *#anjisalvacion*, *#nowplaying*, *#tejran*, *#tejasswiprakash*, *#1*, *#ukraine*, *#bbnaija*, *#90dayfiance*, *#tigraygenocide*, *#treasure*, *#whatshappeninginmyanmar*, *#100daysofcode*, *#bitcoin*, *#writingcommunity*, *#smackdown*, *#maga*, *#wweraw*, *#loveisland*, *#cdnpoli*, For each hashtag 100 samples from 30% of its tweets were drawn and classification via the smallest dissimilarity to the hashtag spectra was performed. The computations were performed with increasing number of hashtags from this list. The results are shown in Table 2. For first two hashtags were taken (*#lolinginlove*, *#puredoctrinesofchrist*), no classification error was made. When the third was included (*#anjisalvacion*), 1.3% error was observed. With 11 hashtags, 9.8% classification error was observed. The F1 measure is also reported in this table.

4. Conclusions

We have elaborated a new characterization of topical groups of objects, like tweets, via a characteristic spectrum of combinatorial Laplacian. It

no. of hashtags	2	3	4	5	6	7	8	9	10	11
error %	0.00	1.33	1.75	0.60	4.83	7.00	9.75	9.22	8.20	9.82
F1*100	100.00	98.67	98.25	99.40	95.18	93.03	89.99	90.53	91.71	90.03
no. of hashtags	12	13	14	15	16	17	18	19	20	21
error %	14.17	16.77	19.07	18.47	20.75	23.53	26.22	30.47	26.90	29.38
F1*100	85.74	83.18	80.86	81.00	78.90	76.04	73.61	69.37	72.89	70.60

Table 2. Classification errors and F1 measure for most distant hashtags.

appears to be quite a stable descriptor of samples from the same population, while discriminating different populations. Potential applications seem be as ingredients in classification and clustering tasks as well as data visualization and hashtag recommendation [4, 5].

It requires further research as to what causes this spectral behavior for similar and different hashtags.

References

[1] Sevi, H., Jonckheere, M., and Kalogeratos, A. Generalized spectral clustering for directed and undirected graphs. *CoRR*, abs/2203.03221, 2022.

[2] Wierzchoń, S. and Kłopotek, M. *Modern Clustering Algorithms*, volume 34 of *Studies in Big Data*. Springer Verlag, 2018.

[3] Schmidt, A. et al. Using spectral clustering of hashtag adoptions to find interest-based communities. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–7. 2018. doi: 10.1109/ICC.2018.8422244.

[4] Jeon, M., Jun, S., and Hwang, E. Hashtag recommendation based on user tweet and hashtag classification on twitter. In *Web-Age Information Management*, pages 325–336. Springer International Publishing, Cham, 2014. ISBN 978-3-319-11538-2.

[5] Gupta, V. and Hewett, R. Unleashing the power of hashtags in tweet analytics with distributed framework on Apache Storm. *CoRR*, abs/1812.01141, 2018.