International Cartographic Association
Association Cartographique Internationale

The typicality measure as a novel tool for normalising geo-social media data

Eva Hauthal [a,*], Sagnik Mukherjee [a], Dirk Burghardt [a]

[a] *Institute of Cartography, Technische Universität Dresden – eva.hauthal@tu-dresden.de, dirk.burghardt@tu-dresden.de, sagnik.mukherjee1@tu-dresden.de*

* Corresponding author

Keywords: geo-social media, location-based social media, normalisation, bias, typicality, information retrieval, geovisual analysis, spatial-temporal analysis

Abstract:

When analysing data from geo-social media, it is crucial to consider their characteristic that, on the one hand, their spatial distribution is not homogeneous but reflects the population distribution and, on the other hand, that a high-frequency occurrence of certain hashtags or emojis not necessarily implies thematic or temporal relevance. In this respect, a normalisation of these data is indispensable in order to be able to determine relative differences, as only these are meaningful. Our work is going to explore in depth the typicality measure introduced for the first time in Hauthal et al. (2021) which provides a novel means to realise this normalisation for geo-social media data.

Typicality is applicable to a specific dataset (hereafter referred to as sub-dataset) that is taken from a general dataset (hereafter referred to as total dataset). The sub-dataset must always be a part of the total dataset. With the help of typicality, hashtags, emojis or similar can be determined as typical in the sub-dataset in comparison to the total dataset. Sub-datasets can be formed in a number of ways, such as temporally (e.g. sub-dataset from a one-year total dataset covers one month), spatially (e.g. sub-dataset from a global total dataset covers one country) or thematically (e.g. sub-dataset covers one topic of all topics contained in the total dataset). Two relative frequencies are included in the calculation of typicality, which are the relative frequency of a for example hashtag in the sub-dataset and the relative frequency of the exact same hashtag in the total dataset.

We are going to demonstrate that typicality has differences but also similarities to existing measures that serve normalisation. In contrast to relative frequency or tf-idf, typicality does not show proportions or relevance on a unipolar scale, but a-/typical occurrence on a bipolar scale. Typicality is most comparable to signed chi-score, which also provides bipolar values called over- and under-representation. This is because both the calculation of typicality and signed chi-score rely on a general reference dataset, whereby in the case of typicality, unlike signed chi-score, this general reference dataset is very strictly defined. This strict definition that the reference dataset must contain the specific sub-dataset under investigation makes typicality values comparable across different total datasets, even though signed chi-score and typicality do return similar results.

Ongoing work focuses on further research into the typicality measure, in particular the exclusion of hashtags or emojis with a very low absolute frequency. For this purpose, a pre-selection of the most frequently used hashtags, emojis or similar is necessary, as the inclusion of only very little used ones can yield high typicality values, even though such hashtags or emojis cannot be considered typical within a dataset, but rather non-significant. The aim is to develop a standardised procedure for determining the threshold value of the pre-selection. This threshold is influenced, among other things, by the size of the total and sub-dataset, but also by the diversity of hashtags or emojis contained. Another question addresses the effect of the extent of the total dataset on the typicality results. In other words, must the entire available dataset always be used as the total dataset or might even smaller spatial or temporal units be more appropriate? Selecting total datasets of various granularities, along with a consistent sub-dataset could help identify local and global hotspots. Further objects of investigation are appropriate visualisation methods for typicality in a spatial context as well as the required compilation time for typicality calculation in comparison to other measures.

References

Hauthal E., Burghardt D. & Dunkel A. (2021): Emojis as Contextual Indicants in Location-Based Social Media Posts. *International Journal of Geo-Information* 10 (6), Special Issue: Social Computing for Geographic Information Science.

https://doi.org/10.5194/ica-abs-5-72-2022