Research Proposal: Flood coverage analysis based on spatial distributions of Flickr messages

Julian Krauth
Heidelberg University
Study course: Geography
j.krauth@uni-heidelberg.de

Keywords

Social-Media, Flickr, Flood, AGI, Data analysis

INTRODUCTION

Flood-related disasters have significant impact on population and infrastructure (most common disaster type globally from 2006 to 2014), therefore further research could contribute to valuable improvements in terms of mitigation and preparation processes for disaster risk management. In recent years, the amount of collected data has increased exponentially which content can be interpreted to create beneficial aids for disaster management. This assignment proposes research with ambient geographic information (AGI) of Flickr, an image hosting platform where community members can share their pictures with the public. With help of this generated content a method could be created which contributes to the public disaster management systems by appending additional information about flood-related areas where other data-sources are not sufficient for comprehensive social or infrastructural actions.

PROBLEM STATEMENT

As consequence of the climate change, natural hazards are occurring with a larger quantity. Other than the amount of occurring disasters such as floods, the increasing population density in inner cities support the vulnerability of inhabitants. While the threat of natural hazards are becoming more relevant, the amount of collected data, especially user-generated content, is increasing significantly but not commonly used for disaster management systems.
With the beginning of Web 2.0 and its user-driven information social aspects became more and more popular within the internet. Social-Media platforms provide a framework of sharing content. For the proposed research the image hosting platform Flickr is used. While Flickr users are spread commonly over developed and emerging countries the additional data could be used to improve the accuracy of existing flood-related disaster management systems as an supplementary approach of data gathering.
This approach of *citizens as sensors* collects vast amounts of spatial data in realtime whereas further filtering methods needs to be applied. For areas with lack of information of sensor infrastructure, web users can provide useful data, volunteered or in ambient form, which adds significant content to otherwise unknown areas in terms of available data. In this case, with research for flood events, a possible scenario could be missing/ failing water level sensors at a waterflow but several Flickr posts of users with flood-related tags in this certain area of interest. Therefore, this additional information can lead disaster managers to further knowledge about existing floods, for example where flooded areas occurred and at which time certain regions were influenced.
In summary, following research questions need to be addressed:

1. To which extent are flood intensities analysable with help of Flickr messages?
2. Can training methods of historic floods improve the accuracy of flood coverage analysis?
3. Is the theoretical approach of Social-Media based flood analysis feasible to realise in practice?
4. Which parameters does a region need to be suited for flood analysis with AGI?

OBJECTIVES

The long term goal of this research is to develop a framework in which potentials from user-generated content can be derived for disaster management systems, more specifically for flood-related natural hazards. Aimed at creating attention to alternative data sources for the prevention of residents vulnerable to disaster events, even more detailed solutions can be derived from this research. The framework could be used from public administrations in order to improve their disaster mitigation/preparation as well as response to occurring flood events. Following sub-objectives are aimed with this study:

- Test the accuracy of Flickr messages in terms of distributions flood events
- Define suitable areas in which this approach is suitable in practice
- Outline applicable training methods for efficiency maximisation

The proposed area of interest has to be affected to a high rate of flood-related events, low quality or quantity of water level measuring infrastructure and a high level usage rate of Social-Media, in this studies case a high usage of Flickr.

(welche Datengrundlagen, woher beziehbar, zugänglichkeit, Eignung, Methoden zur Datenanalyse rechtfertigung

PRELIMINARY LITERATURE REVIEW

METHODOLOGY

To conceive qualitatively outcome of the research a possible approach in advance is to limit the input source messages of Flickr. With more than 3.5 million new images uploaded every day **(The Verge**; **20 March 2013)** data collecting without filtering them beforehand is resource intensive in terms of data storage and temporal aspects. Therefore, filter methods which limit the area of interest are needed in order to improve further data processing and analysing.

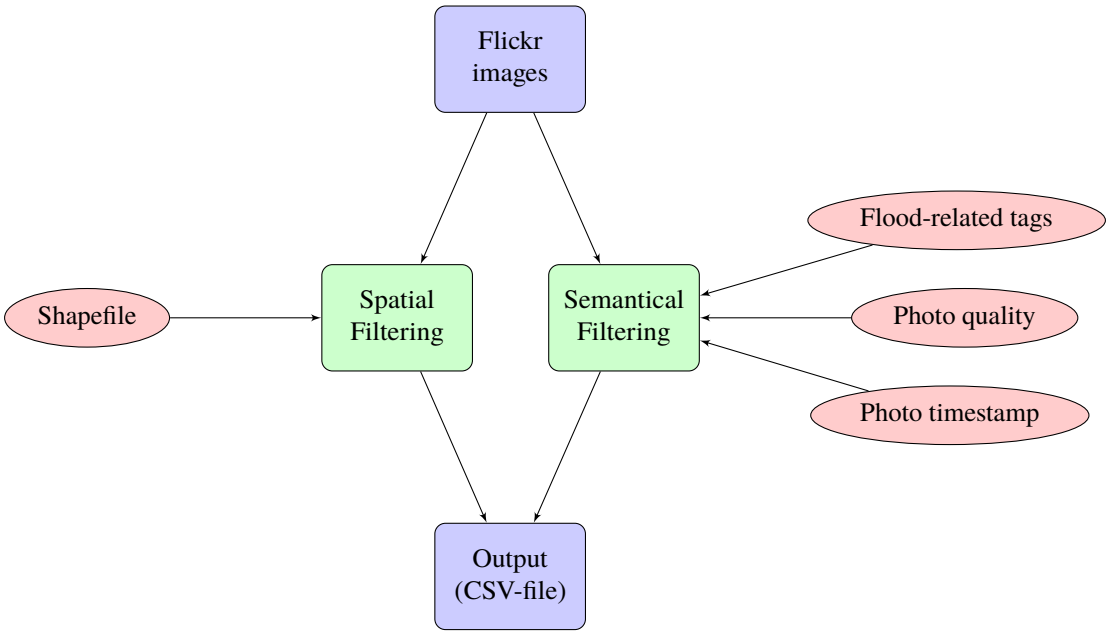

Figure 1. Worflow-diagram for usage of PythonAPI

Spatial Filtering

With the knowledge that floods occur mostly along rivers or coastlines, first spatial limitations are obvious. Further, more detailed restrictions can be derived by the actual research area where the flood analysis takes place. If a flood needs to be analysed which area is concentrated to the surrounding of a certain water flow, spatial filtering can guarantee that the Social-Media images are shot within this specified zone.
In practice, the Flickr Python Api can be manipulated in a way that custom Latitude/Longitude boundaries define where the data is collected. This specific filtering method creates a rectangular box with the individual boundaries where all the Flickr pictures shot within are selected and every image outside the box is dropped. Nevertheless, geographical forms are rarely linear and effectively filtered with a rectangle. With small adjustments to the Api-code, other with other shapes can be filtered too.
In this code-snippet, a more accurate spatial filtering method has been applied. To achieve increased precision in terms of spatial filtering, the user can load a polygon shapefile to the script which is then automatically read. In detail, the python script reads the input shapefile with a polygon of the research area of interest. Following to reading the file, the specific latitude/longitude coordinates of peak values are automatically returned. In this case, the minimum longitude, minimum latitude, maximum longitude and maximum latitude create a bounding box in which the FlickrAPI search request is called. Basically, the API is searching in a box with four nodes, whereas a polygon can relate to a much more detailed shape. Later in calculation process, the results are spatially filtered to the exact extent of the polygon in the shapefile (here from line 18 to 22).

```
1 [...]
2 try:
3     infile_shape = sys.argv[1]
4     outfile_csv = sys.argv[2]
5
6 except:
7         print "USAGE of this script: Shapefile_of_Bounding
8         _Box.shp Output.csv"
```

```
9           sys.exit()
10
11  drv = ogr.GetDriverByName('ESRI Shapefile')
12  ds_in = drv.Open(infile_shape)
13  lyr_in = ds_in.GetLayer(0)
14  extent = lyr_in.GetExtent()
15
16  [...]
17
18  #set spatialfilter --> shapefile vs point
19  lyr_in.SetSpatialFilter(point)
20  #if point inside shapefile
21  for feat_in in lyr_in:
22          [...]
```

The spatial filtering process is created with help of the GDAL/OGR Python API, in this example, a spatial filter is set on the input shapefile layer with coordinates of examined points (line 19). After setting a spatial filter, the requested points are checked whether or not the coordinates of the image metadata are spatial aligned with the exact boundary of the input shapefile polygon.

Semantical Filtering

Photo Timestamp The source data of Flickr images require not only spatial correspondence to the researched topic, additionally semantic filters have to be applied to the input data, otherwise random photographs from the actual research area appear in the output. Based on the subject that needs to be researched, different filters are required. In the proposed example of flood analysis, photo timestamps have to be considered, basically from which time period images are accepted for further research. Floods occur often in specific time frames, mostly from hours to several days. It is expectable if a certain flood is researched, the time interval is set to length of the natural hazard. For longterm flood research analysis, more extended intervals have to be set.
An advantage of Flickr (and also a reason why this Social-Media platform is used) is its image metadata storing. Besides camera type, shutter speed and exposure time of a taken image, also the date uploaded and date taken is stored and published. This feature leads to feasible approaches in practice, where the time of flood occurring is well known, and therefore specificly selectable via the photo timestamp filter option. The API provides a time filter option within the photo-search dialog with which the upload-date or the date when an image was taken can be specified:

```
1  photo_list = flickr.photos.search
2  (min_taken_date='2007-01-01 00:00:01')
```

In this projects example, the filter is set to gather information from pictures that are taken after January, the first in 2007 at 00.00 o´clock and one second.
Image Quality A different aspect of semantical filtering relates to qualitative suitability of gathered images. Flickr classifies the uploaded images in different quality types. This classification reaches from original, uncompressed photos of random size to several large pictures (1024 to 2048 pixels on longest side), medium sized pictures (500 to 800 pixels on largest side) and small pictures (75 to 320 pixels on largest side). As evaluation, this feature is required if visual image analysis needs to be performed or minimum standards are expected for research purposes. Flood distribution analysis has its main focus on the allocation of specific images inside the research area, thus the image quality has a subordinate role. An appropriate quality is only required for conclusive manual optical filtering further described in the next chapter.

Topic filtering Without definition of flood-related tags simply every picture without specific context would be returned, whether or not a flood is shown. The creation of suitable tags is dependant on thematic impressions and locational extent. Thematic parameters are required in case of individual search focus, in this proposed research flood designations. The locational aspect is relevant in terms of lingual expressions of Flickr users. For example, a research area in Germany needs other tags than a region in Australia, some expressions have to be in German, e.g. *Flut* or *Überschwemmung*. The most used foreign language in Germany is English, so the lingual expressions for flood should be included as well in order to create a sufficient output. These parameters have to be considered and manipulated in case of differences in geographic expansion. As example, a case-study in Germany could be performed with following parameters as tag-search:

- Flood
- Floodplain
- Flut
- Überschwemmung
- Hochwasser

As explained above, the tags are chosen by the most common description of flood events in the researched area of interest, for areas in France, the German expressions have to be replaced with their French equivalents. Beware, searching for multiple tag expressions increases the possibility of redundant output data. It is imaginable that a photo of a floodplain is tagged with: *flood*, *floodplain* and *Hochwasser*. This example would create an equal output three times. To prevent this redundancy from occurring, a simple **set** can be generated, where the specific URL of the image is stored:

```
1 cnt_final = 0
2 uniqueurl = set()
3 if url not in uniqueurl:
4   cnt_final +=1
5   outFile.write(add)
6   uniqueurl.add(url)
```

At first, a counter is initiated which counts the final amount of photos that are written in the Output-csv file. For using only distinct pictures, a new empty set is created (see line 2). The advantage of set in contrast to a list is the storage of distinct values only. Even if more than one equal value has been tried to store in a set, only one of it is listed. The if-condition afterwards checks whether a URL of an image is already in the set or not. If it is, no further action is necessary because the same picture has already been stored. For a true condition, the counter adds one to its value, the image properties are written in the output-csv file and the photo URL is stored in the set() in order to prevent a duplication.

Postprocessing

The final output of the filtered messages are formatted in a CSV-file which allows list structures with header instance. As shown in the table, the header is filled with fid, uid, characterization tags, recording date, latitude, longitude and specific URL link to image (accessible in web-browser). As seen in column 3 (tags), the prevention of redundancy is necessary. Multiple images have more than one acceptable keyword as tag (acceptable keywords for this case-study: *flood, flut, überschwemmung, hochwasser, floodplain*).

fid	uid	tags	date	lat	lon	url
101813	12832@N05	flood heidelberg hochwasser	2013-05-02 22:45:12	49.4151	8.7071	https://..
018712	43122@N02	hochwasser neckar brücke	2014-02-12 12:35:52	49.4153	8.7081	https://..
...	...	...	...	...	...	...

In case of no efficient redundancy prevention function, several images would be duplicated inside the output file. The first two columns define the ID of the image itself and the ID of the owner who uploaded the picture. For a more detailed view of the outcome quality manual visual analysis have to be added. Depending on the use-case false positives should be deleted from the final dataset, for example different meanings of the tag expression. Possible false positive could be Flickr posts with the tag flood or überschwemmung after a water pipe rupture inside the clients resident. In order to avoid distorted statistics these messages have to be manually filtered out.
To analyse the distribution of every positive outcome message, the CSV-file can be imported in a Geographic Information System (GIS). Some GIS, for example QGIS, support direct CSV import where only Latitude and Longitude values have to be present in the table. The Flickr API stores this information as metadata and therefore, is listed in seperated columns in the table. In QGIS, each Flickr image is displayed as point data which only have to be set as input for the QGIS *convex hull*-function (QGIS Documentation, http://docs.qgis.org/2.18/pdf/de/). If the usage of a GIS is undesired, inserting the data in a database is an alternative way to go. For this specific purpose, PostGIS, the spatial database extender for PostgreSQL, has a similar function called *ST_ConvexHull* which creates a geometry that represents the minimum convex geometry that encloses all geometrie within the set

(PostGIS Documentation, https://postgis.net/docs/ST_ConvexHull.html). With the resulting geometry/polygon, the flood-related messages are spatial aligned to an area where the defined time interval fits to the requested image quality and search image tags. This area displays the analysed flood from the Social-Media side of view which then could be further analysed/compared to other flood coverage models.

Comparing analysis to existing flood analysis models

The concluding analysis method addresses the comparison between existing flood distribution analysis models and the just computed Social-Media approach. How high is the accuracy from Flickr image analysis as standalone method as well as addition to previous monitored systems? To which extent does the precision change depending on time (user count and user activity rate as crucial factors), on area of interest (user activity in certain regions as crucial factor) and on infrastructure (demographic density, power/internet availability as crucial factors)?
These two questions are raised by the research proposal as well as the third and last research question which training methods could be applied for efficiency maximisation, can be answered by executing the above described methodology with a following comparison of different study-areas, results of Flickr-only approach to existing water-level measurement systems as well as historic flood data where findings for further research could be gathered.