paper:5

PhoCA: An extensible service-oriented tool for Photo Clustering Analysis

Yuri A. Lacerda[1,2], Johny M. da Silva[2], Leandro B. Marinho[1], Cláudio de S. Baptista[1]

[1] Laboratório de Sistemas de Informação
Federal University of Campina Grande (UFCG) Campina Grande, PB – Brazil

[2] Federal Institute of Education, Science and Technology of Ceará (IFCE)
Crato, CE – Brazil

{yurilacerda,johny.moreira}@ifce.edu.br,
{lbmarinho,baptista}@dsc.ufcg.edu.br

Abstract. *Clustering algorithms are at the core of most of the research work in knowledge discovery using geo-tagged photos, e.g., a cluster of photos might identify existing or new points (or areas) of interest. But before clustering the data, researchers need to devote a lot of time and effort for collecting, pre-processing, and analyzing the data. For assisting researchers in these tasks, we propose PhoCA[1] (Photo Clustering Analyzer), an extensible web-based tool for photo clustering analysis. It aims at assisting the researcher in tasks such as data collection, clustering analysis, and visualization. PhoCA features tools for collecting and extracting metadata of geo-referenced photos; suites of state-of-the-art photo clustering algorithms; spatial visualization tools; etc.*

1. Introduction

The current generation of domestic and professional digital cameras enables the production of high quality photos enhanced with rich contextual information, such as the date, time, and location of the photo at shooting time, as well as the direction of the camera in the moment of shooting [Lacerda et al. 2012].

The large amount of photos available in the Web, combined with their rich contextual metadata, opens new opportunities for knowledge discovery and data mining applications. Some recent areas of research worth mentioning concern the automatic organization of personal photo collections [Figueirêdo et al. 2012]; the automatic planning of touristic routes [Popescu and Grefenstette 2009]; the automatic discovery of points of interest (POI) [Lacerda et al. 2012]; etc.

In all these applications, clustering algorithms appear as one of the main tools. For example, most of the existing research on detection of points of interest from digital photos is based on new or modified versions of existing clustering algorithms [Lacerda et al. 2012][Yang et al. 2011][Kisilevich et al. 2010]. In the area of recommender systems, the authors of [Matyas and Schlieder 2009][Marinho et al. 2012] propose to first identify areas of interest through clustering of geo-referenced photos, and then recommend the detected areas of interest to the users.

[1] A screencast and the PhoCA installation files are available on: http://www.yurilacerda.com/phoca.

However, before one can finally run clustering algorithms for investigating the aforementioned applications, one needs to first collect and pre-process the data, which is very time consuming, especially for new researchers in the area. This involves studying the data access API of the photo repository of interest and writing specific scripts for collecting the photos and extracting the metadata of interest. Thereby, it would be of great value to have applications that could assist the researchers in these tasks.

In order to fulfill this need, we propose PhoCA (Photo Clustering Analyzer) an extensible web-based tool for photo clustering analysis. PhoCA aims at assisting researchers in this field through the most important tasks of the KDD process, such as data collection, pattern recognition, analysis, and visualization.

This work is organized as follows. In the next section, the related work is presented. Section 3 presents the architecture and describes the modules that comprise PhoCA. Section 4 presents a case study using PhoCA. Finally, the Section 5 concludes the paper and discusses opportunities for future work.

2. Related Work

Although there are many works about using clustering algorithms for different applications, to the best of our knowledge PhoCA is the first specialized tool for assisting researchers in most of the steps of the KDD process for this particular domain. Some of the most popular problems in this domain are: the point of interest detection ; identification of representative images of landmarks; and automatic planning of touristic routes.

Regarding points of interest detection, most of the approaches cluster the photos based on their geographic distances. The assumption is that many photos taken close to each other tend to indicate points of interest. The most common algorithms found in the literature for this end are: k-means, mean shift, DBSCAN [Kriegel et al. 2012], and spectral clustering [Yang et al. 2011].

Some studies use DBSCAN to cluster geo-referenced photographs [Lacerda et al. 2012][Kisilevich et al. 2010], however, the original DBSCAN algorithm does not allow clusters with density variations [Han et al. 2011]. In [Kisilevich et al. 2010], it is proposed P-DBSCAN, a new approach that uses the DBSCAN algorithm, taking into consideration just a local density, that is, allowing the creation of clusters with different sizes and densities.

In [Lacerda et al. 2012] we proposed a new clustering algorithm that uses photo orientation and geotags of photographs to discover new points of interest. The algorithm creates line segments using photo orientation and uses the points of interceptions of these lines and some photo positions as input for the clustering process. All these algorithms are available in PhoCA.

In the reviewed literature, we did not find any tool that is similar to PhoCA. Instead, most of the data mining tools available are of general purpose, such as Weka [Hall et al. 2009].

However, differently from them, PhoCA handles the specific task of clustering large collections of photos. Thus, most of the functionalities of PhoCA are not featured by the aforementioned tools.

3. System Architecture

Figure 1 presents the architecture of PhoCA. The system has a service-oriented architecture, which means that its functionalities are available through the invocation of Web services. This enables the access of the functionalities of PhoCA from other software, independent of programming languages and system platform.

PhoCA is a RESTful application. The user interface was developed using HTML (Hypertext Markup Language), CSS (Cascading Style Sheets) and Javascript. Also, the JQuery Javascript Library was used to enable rapid Web development, handling events and Ajax interactions with the services.

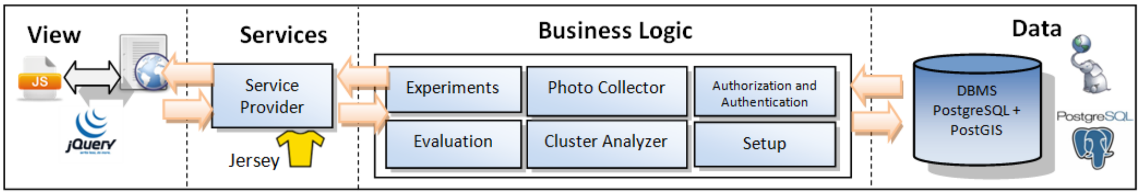


Figure 1. PhoCA Architecture

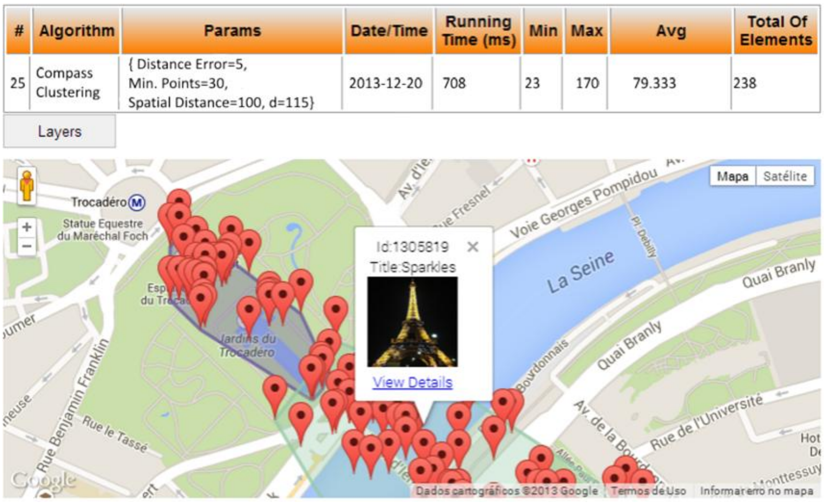
The model layer has the java classes responsible for the business logic. It divides the system into six modules: Photo Collector; Experiments; Cluster Analyzer; Evaluation; Authorization and Authentication; and Setup.

The first module is responsible for the extraction of metadata from local photos or Flickr. The system allows the storage and metadata extraction of photos uploaded by the users. In case of photos from Flickr, the user must inform a specific bounding box area to extract photo metadata and a keyword. Depending on the chosen parameters, in especial the bounding box size, this process could be expensive, e.g., taking many hours or even days. Therefore, there are some parameters (eg: number of threads used on the photo collector) on the Setup Module that must be tuned for performance.

The Experiments module allows the execution of a clustering algorithm for a specified collection. PhoCA provides five clustering algorithms, however the user could easily add new algorithms using some object-oriented programming elements and adding some lines in a XML file.

The next module, Cluster Analyzer, is responsible for presenting summary statistics related to the computed clusters. Beside this, the Evaluation module allows the users to evaluate if a specific photo is correctly associated to a group using a Graphical User Interface (GUI). Moreover, this module calculates the metric of precision based on the percentage of photos clustered correctly. This is a common metric in information retrieval field.

The Summary Statistics module is responsible for presenting simple statistics for a specified experiment, such as: the experiment identification, cluster method and parameters used; data and time of execution; running time (in milliseconds); and the mean, minimum and maximum cluster size. Moreover, this module is in charge of presenting the size of each cluster, allowing the visualization of these elements in a map, Figure 2.



#	Algorithm	Params	Date/Time	Running Time (ms)	Min	Max	Avg	Total Of Elements
25	Compass Clustering	{ Distance Error=5, Min. Points=30, Spatial Distance=100, d=115}	2013-12-20	708	23	170	79.333	238

Figure 2. The Summary Statistic Module

The Authorization and Authentication module is in charge of the system's security, controlling to resources and functionalities. Furthermore, it allows the creation and manipulation of system users. There are two user profiles: administrator and default user. Only the administrator profile has privileges to create new users and changing the system's settings. The default user profile has privileges to use the other modules through system's GUI or invoking its Web services from other application.

The Setup module is responsible for setting the database access, Flickr API Key and performance tunings. PhoCA must be installed on a Web container. A Web container is a component of a Web server that implements the Java Servlet specification. Otherwise, it is not possible to access its functionalities. PhoCA uses as database management systems the PostgreSQL[2] and the spatial extension Postgis[3].

4. Case Study

In this section, we present a case study in order to evaluate the proposed tool. This case study runs over all the knowledge discovery process, i.e., it starts with the collection and storage of photos, and culminates with the execution of experiments using the Compass Clustering algorithm proposed in [Lacerda et al. 2012] for landmark detection.

We consider a photo related to a landmark as a photo captured inside the landmark area or that contains the landmark in the image content. The idea here is basically to check which size of the line segment, one the parameters required by Compass Clustering, provides the best results for each data set.

It was extracted three photo collection of a bounding box area around 500 meters of 3 well-known POIs: Eiffel Tower (Paris, France), Statue of Liberty (New York, USA) and Colosseum (Rome, Italy). We extracted metadata from 36,230 geo-tagged photos from which 1,109 are oriented. These photos were filtered manually to discard invalid landscape photos. Finally, we used 704 valid landscape photos in our evaluation: Eiffel Tower (316 photos), Colosseum (277 photos) and Statue of Liberty (111 photos). These

[2] http://www.postgres.org
[3] http://www.postgis.org

collections had 71,51%, 85,92% and 84,68% of its photos related to respective landmark. The valid landscape photos contain correct data about orientation and geo-location and they haven't focus in a specific object. We made a manual inspection for each photo.

We executed the experiments using the Compass Clustering algorithms with the line segmentation size (d) varying between 20 and 500 meters with steps of 20 meters. It was used the DBSCAN algorithm as the internal clustering algorithm of Compass Clustering. We used the parameters minimum points = 30 and spatial distance = 100 meters. These values were inferred by subjective observation. We set the constant distance error = 5 meters. This constant does not consider the interception points of photos that are less than the distance error. This was motivated because of GPS imprecision of cameras and smartphones GPS chips.

We analyzed the metrics precision and recall. Precision was defined as the ratio between the total of photos related to landmark in the cluster and the total of elements of the cluster itself. Recall is computed as the ratio between the number of photos related landmark in the cluster and the total number of photos related landmark.

Figure 3 presents the average of precision p and recall r of all clusters for the three collections: Eiffel Tower (p1 and r1), Colloseum (p2 and r2) and Statue of Liberty (p3 and r3). The precision and recall have become almost constant to values of d > 360 meters. In other words, the increasing of d hasn't difference to values higher.

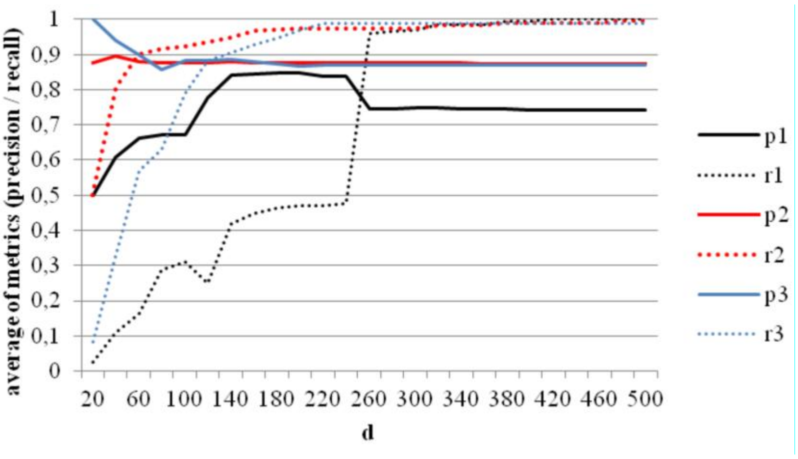


Figure 3. Precision and Recall of Experiments

5. Conclusion and Future Works

This paper presented PhoCA, an extensible web-based tool for photo clustering analysis. This tool features: public access through web services; extracting metadata of photos from Flickr or local photo repositories; clustering photos using state-of-art algorithms; and the analysis of clusters through a geo-spatial visualization tool.

Among the reviewed related work, none of them proposes a specific tool for photo clustering analysis, including tools for all the steps of the KDD process. We presented an experiment for landmark detection using the algorithm Compass Custering. The algorithm presented the influence of the parameter line segmentation.

Future work include to extend the suite of clustering algorithms available with other algorithms recently used for knowledge discovery in large photo collections (e.g., mean-shift) and to perform comparisons between several algorithms. We also plan to enhance the Cluster Analyzer module with more sophisticated filters, such as the visualization of the trajectory of the user while shooting photos, and the visualization of photos that were shot within a given period of time.

6. Acknowledgment

The authors are grateful to the CNPq and IFCE for funding this research project.

7. References

Figueirêdo, H. F., Lacerda, Y. A., Paiva, A. C., Casanova, M. A. and Baptista, C. S. (2012). PhotoGeo: a photo digital library with spatial-temporal support and self-annotation. *Multimedia Tools and Applications*, 59 (1): 279-305.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.

Han, J., Kamber, M. and Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier, 3th edition.

Kisilevich, S., Mansmman, F., and Keim, D.A. (2010) P-dbscan: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In *Proceedings of COM.Geo'10.*

Kriegel, H.-P., Sander, J., Ester, M. and Xu. X (2012) A density-based algorithm for discovering clusters in large spatial database with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226-231.

Lacerda, Y. A., Feitosa, R., Esmeraldo, G. A., Baptista, C. de S. and Marinho, L. B. (2012) Compass clustering: a new clustering method for detection of points of interest using personal collections of georeferenced and oriented photographs. In *Proceedings of the Webmedia '12*. pages 281-288.

Marinho, L. B., Sandholm, T., Baptista, C. de S., Nunes, I., Nóbrega, C. and Araújo, J. (2012) Extracting Geospatial Preferences Using Relational Neighbors. In *Journal of Information and Data Management (JIDM)*, 3 (3): 364-377.

Matyas, C. and Schlieder, C. (2009) A spatial user similarity measure for geographic recommender systems. In *Proceedings of Third International Conference on GeoSpatial Semantics*, vol. 5892 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pages 122–139.

Popescu, A.; Grefenstette, G. (2009) Deducing trip related information from flickr. In *Proceedings of WWW'09*, pages 1183–1184.

Yang, Y., Gong, Z. and Hou, L. (2011) Identifying Points of Interest by Self-Tuning Clustering. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR'11)*, Pequim, China.