In partial fulfillment of the terms for obtaining the PhD degree, Suela Isaj will give a lecture on the following subject:

Multi-Source Spatial Entity Extraction and Linkage

on Tuesday 8th of June 2021, 13:00,

Abstract:

Web data sources contain large amounts of geo-social data, consisting of users, friendship/follower networks, check-ins, reviews, locations, etc., which are of great interest to academia and industry. There are publicly available datasets of samples of this web data, but they are very old (over ten years), not large, and not rich in attributes. Alternatively, one could use the public APIs to access and download web data. Unfortunately, this process is challenging due to the APIs limitations (e.g. the amount of data retrieved in a request, the number of requests performed within a timeframe, etc.). Thus, there is a need for algorithms that, given the limitations, are able to retrieve a good quality dataset from web sources, a need that the current state-of-the-art does not address.

This thesis aims to provide algorithms and tools that can produce larger, recent, duplicate-free, and rich-in-attributes spatial entity data. To obtain larger and recent data from web data sources, we propose multi-source seed-driven (MSSD) algorithms that use the public free APIs to extract geo-social data. The MSSD algorithms aim to maximize the amount of data extracted while minimizing the number of requests and respecting each source's API limitations. The rationale behind the seed-driven algorithms is to perform some API requests for having an initial dataset and then use the points of the richest source as seed in the API requests for the rest of the sources. We propose different techniques for choosing the points and the radius of the search. We opt for a multi-source solution given that multiple sources provide independent information and diverse attributes as opposed to using only one source. Moreover, we experimentally demonstrate that using a single source algorithm sometimes converges to a dead end. The MSSD algorithms extract overall 14.3 times more data than the initial querying, and the optimized version MSSD* retrieves 90% of the data with less than 16% of the requests of the non-optimized version.

When obtaining multi-source data, the same spatial entity might exist in different sources and sometimes even within the same source. These "duplicates" are not easy to detect since they have different attributes, they are expressed in different forms, and they might even contain contradicting attribute values. The problem of finding which pairs of spatial entities refer to a real physical entity is referred to as *spatial entity linkage*. We address this problem with several algorithms, which all share the same spatial blocking technique, and they use skylines to rank the compared pairs. The spatial blocking technique (QuadFlex) that we propose is a quadtree-inspired algorithm that groups the spatial entities based on the distance between them and the area's density. Moreover, it allows the assignment of spatial entities in more than one child to not miss any relevant comparisons. The spatial entities that fall into the same child are compared pairwise.

To decide which pairs belong to the same physical entity, we propose novel skyline-based (SkyEx-*) algorithms, which use preference functions to assign skylines to the pairs. The threshold-based SkyEx, SkyEx-F and SkyEx-FES require a threshold that is the number of skylines to separate the positive from the negative class, and they are able to achieve an F-measure of 0.72 on the whole dataset and 0.85 on a manually labeled sample. We introduce a fully unsupervised algorithm, SkyEx-D, which does not need a threshold and instead sets the cut-off based on the distance of the skylines. We demonstrate experimentally that SkyEx-D can reach a near-optimal F-measure (less than 0.01 loss). Additionally, we offer `skyex`, an R-package that implements the threshold-based and unsupervised skyline-based algorithms, supports the whole entity linkage pipeline with other state-of-the-art methods for entity blocking and comparisons, and provides a powerful Analysis and Visualization module to aid the explainability of the results.

Besides the unsupervised algorithms, we propose a trained skyline-based algorithm, SkyEx-T, which is able to learn the preference function and the cut-off in tiny training sets (0.05%-1% of the dataset) and still achieve machine-learning-level accuracy. Moreover, the SkyEx-T model is fully explainable and readable, in contrast to the commonly-used black-box machine learning techniques. Furthermore, SkyEx-T has no weights nor layered architecture; consequently, it shows high robustness in deployment, while for the machine learning, some re-configuration and re-tuning of parameters might be needed when the new data arrives. Finally, we demonstrate that SkyEx-T cut-off closely approximates the optimal cut-off, even though it was learned on a tiny training set. With our algorithms in the spatial entity linkage, we ensure a duplicate-free dataset and rich-in-attribute spatial entities.

Overall, this thesis contributes with effective and efficient algorithms for the initial and fundamental step of every geo-social research study: having recent, good-quality, rich-in-attributes datasets. We propose the MSSD-* algorithms that make the data extraction process more effective (14.3 times more data than the initial querying) while managing the requests carefully. We further improve the quality of the retrieved data by detecting pairs that refer to the same entity with high precision and recall while having an explainable and robust model (SkyEx-* algorithms). In the future, in the context of data extraction, we aim to work on hybrid algorithms that combine location-based with user-based and keyword-based API requests and use supervised techniques to learn the parameters of APIs. In the context of spatial entity linkage, we plan to work on hybrid blocking techniques that combine spatial attributes with textual and semantic ones, multi-class classification for the skyline-based algorithms, and crowdsourcing techniques for improving the labeling of the pairs.

Members of the assessment committee are Associate Professor Manfred Jaeger, Aalborg University, Denmark, Associate Professor Maria Luisa Damiani, University of Milano, Italy, and Professor Konstantinos Stefanidis, University of Tampere, Finland. Professor Torben Bach Pedersen and Professor Esteban Zimányi are Suela Isaj's supervisors. The moderator is Associate Professor Christian Thomsen.