Active Learning Approach to Record Linking in Large Geodatasets

Alexandre Sorokine [a]*, Jason Kaufman [a], Jesse Piburn [a], Robert Stewart [a]

[a] Oak Ridge National Laboratory, Oak Ridge, TN, U.S.A.
* SorokinA@ornl.gov

Keywords: dataset conflation, geographic object identity, Big Data

Introduction

Data integration from different sources is a common challenge in the present-day geoprocessing. One of the tasks that has to be performed in many integrations is finding correspondence between the features of different geodatasets. In literature this task or its various aspects can be called conflation, data matching, record linking, entity resolution or alignment. The case when records referring to the same real-world object are detected is of special interest because it is used in many typical integration scenarios like merging datasets, updating or cross-validating records, deduplication, and others. As a result of the ever-growing volume of data to be conflated, manual matching is becoming more and more costly. The need for effective conflation algorithms is likely to become more pronounced in the future as the level of automation of the data processing operation will increase leaving little space for human involvement. Here we evaluate challenges of record linking in the current data-rich geoprocessing environments and propose an approach based on Machine Learning (ML).

Problem Statement

The goal of conflation or matching process can be either a new dataset that incorporates original data in part or as a whole. Conflation can be used for other purposes like cross-verification of the datasets, filling the gaps, updating with newly acquired records, establishing sameness, or other types of relations among the features. Typical data matching and record linking workflow is shown on Fig. 1. At the preprocessing stage the data has to be converted into a common format or accessed through a programming interface. At the next step all records are compared pairwise and using some similarity metrics are classified into matches, possible matches and none-matches. Finally, matches are evaluated for correctness and some of them may be reconsidered.

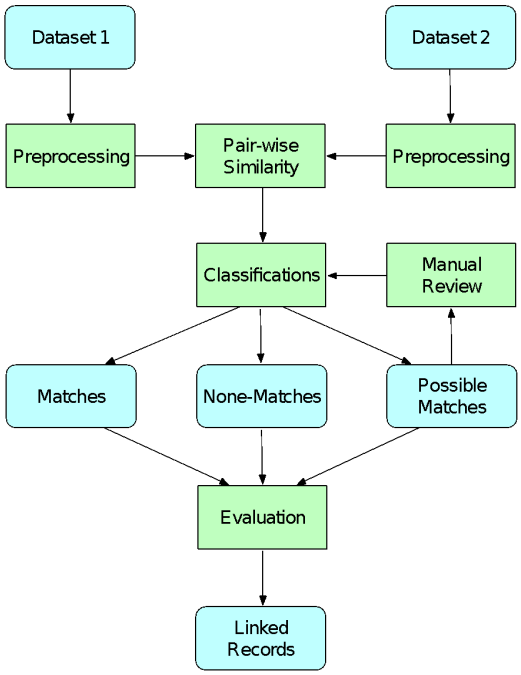

Figure 1Record Linking Workflow (based on Fig. 21, page 24 in Christen, 2012)

Today the main challenges of data conflation are related to the Big Data revolution that concerns growing volume and variety of the data (Karimi 2014). It is no longer uncommon to have the datasets in excess of dozens of millions of records. Such volumes of data demand higher levels of automation. In terms of variety the majority of large datasets have incompatible or loosely defined schemata and data curation processes. Matching criteria for such features cannot be hardcoded and require human interaction. In this study we are developing a highly automated geodata conflation method that is suitable for processing of large volumes of diverse data (millions of records and larger). The method is able to automatically deduce and improve matching criteria based on the input from human experts and has the built-in capability to evaluate the efficiency and quality of the matching results.

Earlier Work

Automated map conflation has been discussed at AutoCarto conferences since at least 1985 and the term "conflation" has been in use even earlier (Lynch and Saalfeld 1985). For a recent review see, for example, (Sun, Zhu, and Song 2019). Outside of the geographic domain, data matching is commonly used for medical and census records, bibliographies, product catalogs, inventories, etc. with early examples going back as early as the 1960s (Christen 2012; Fellegi and Sunter 1969). In early works on geodata conflation much attention has been paid to geometric alignment of the features (Lynch and Saalfeld 1985). The majority of the present interest is in the area of Volunteered Geographic Information (VGI) (*ngageoint/hootenanny* [2015] 2020). VGI supplies enormous amounts of very diverse but also very impure data. Both problems (volume and diversity) can be addressed by application of the ML methods. ML may reduce the need for hard-coding feature similarity measures (Winkler 2002). Expert knowledge can be utilized with the help of active learning that is an ML methodology that relies on user

feedback to improve the model (Sarawagi and Bhamidipaty 2002; Sarawagi et al. 2002). Active learning has been experimented with in data matching outside of the geographic domain (Arasu, Götz, and Kaushik 2010). However, the results of these studies are not directly applicable to geodataset and we want to fill this gap.

Analysis of the Challenges

There are many reasons why a singular entity may have multiple records in different or even the same datasets. In medical and census data, records are often entered separately lacking a common identifier. In that case finding matching records can be done by comparing salient discount errors, spelling variations, missing values or detection of special circumstances like change of name or gender. In general, in the medical or social context identity of a person behind a record is rather clear and easy to understand. However, in our domain the concept of a geographic object identity needs clarification.

Locational information associated with the records is the hallmark of geodata that makes it different from other domains. All features in geodatsets are georeferenced and this information can be used to significantly reduce the number of potential matches across conflated data. It is safe to assume that nearby or overlapping features would be at least related to each other or may even represent the same real-world object. However, different levels of positional accuracy of the datasets may require additional effort to match the records as multiple match candidates may fall into error bounds. They may or may not represent the same real-world entities. The other problem specific to geodata is the geographic scale. The same feature may be represented differently at different scales and finding the proper match typically is not trivial.

Matching of the records collected at different times is a challenge both in and outside the geodata domain. The real-world objects may change or move while the records about them also evolve. This is an especially hard problem for man-made features that can be constructed, repurposed, or cease to exist. For example, it is not clear if a bridge and a disused bridge in the same location should be treated as the same object. This example also demonstrates schema-level incompatibilities of the datasets that may use conflicting or incomplete feature definitions (Feng and Sorokine 2014). Another common problem is the case when conflated features are related to each other (*e.g.*, an entrance to a store and the store itself) but this relation cannot be expressed in the dataset schema.

Data

To demonstrate the applicability of the proposed method we use the Digital Nautical Chart (DNC®) dataset from the National Geospatial-Intelligence Agency ("Digital Nautical Chart" 2020). The U.S. portion of this dataset can be freely downloaded from the agency website. DNC® is organized in 4 scale levels but the same features in different scale levels are not linked to each other. We use our method to link the records across the scale levels. The main challenges here are (1) the size of the data (more than 4 million features) that makes manual conflation very costly, (2) requirement for highly reliable results, (3) significant temporal span of the data collection events, and (4) issues related to multiple representations at different geographic scales.

Proposed Approach

All DNC records have been loaded into a single PostGIS table. The list of match candidates was created using minimal Euclidean distance between features within a feature class. Feature pairs separated by distance exceeding a predefined threshold were eliminated. The pairs with identical attributes were considered matched and were excluded from further processing. For each remaining pair we have calculated a similarity vector $S_{i,j} = [d_1, d_2, \dots, a_1, a_2, \dots]$. Similarity vector was designed to take multiple parameters into account and can be adjusted to the needs of specific feature classes or extended. To evaluate geographic proximity in addition to minimal Euclidean distance we use Hausdorff and Fréchet distance and percentage of the buffered overlap between feature geometries. Attribute similarity calculation depends upon the type of the attribute. For physical measurements we use value difference normalized on standard deviation. For categorical values we use a Boolean flag that indicates an exact match. For text fields that represent entity names we use Levenshtein distance ("Levenshtein Distance" 2019). The sets of attributes are compared using Jaccard coefficients. Our similarity vector can be extended to account for neighborhood information, *i.e.*, presence of features of the same or other classes in the surrounding areas. Also, it is possible to use more advanced comparison of the text attribute based on the natural language processing methods.

The system is intended to be implemented as a recommender for an expert who performs conflation. For each feature the user is offered a list of potential matches ranked by the probability of a good match. The probability of the matches is calculated as a linear combination of the components of the similarity vector and predefined weights:

$$Score = \left[d_1, d_2 \dots, a_1, a_2, \dots\right] \cdot [w_{d1}^0\ w_{d2}^0\ \dots\ w_{a1}^0\ \ w_{a2}^0\ \dots\ ]$$

The user chooses the best matching feature based on his judgement and this choice is fed back into the recommender system (Fig. 2). The recommender system adjusts the weights using hierarchical Bayesian logistic regression (Gelman and Hill 2006). The initial implementation is a relatively simple hierarchical model with partial pooling of distance features across feature codes. As a result, the dimensions of the similarity vector that are better at prediction of the sameness of the objects are emphasized while others are played down. The expected result is that the quality of match prediction is improved with user input.

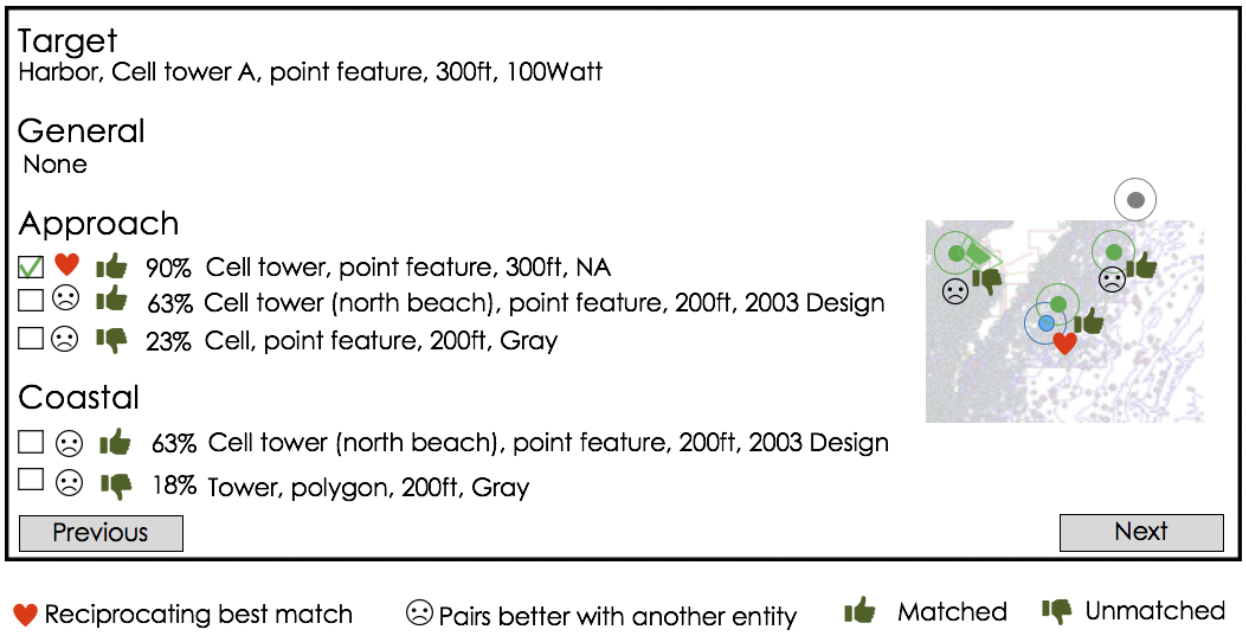

Figure 2Recommender System User Interface Mockup

Conclusions and Future Work

In this study we are evaluating active learning as an approach to record linking in a large multiscale geodataset. Compared to the existing approaches we propose a more general framework to accommodate the semantics of data matching with the ability to integrate existing methods. We use a recommender system with an active learning component to evaluate and improve matching outcomes. Our preliminary results indicate applicability of this approach for our case study and a potential for benefits in terms of automation, utilization of user expertise, and improving the quality of conflation. This is a work in progress and our immediate plans include testing of the approach in the real-world setting. Other improvements would be related to support of special matching cases and optimization of the similarity vector. In the future this approach will become more heavily based on artificial intelligence technologies like deep learning and natural language processing.

References

Arasu, Arvind, Michaela Götz, and Raghav Kaushik. 2010. "On Active Learning of Record Matching Packages." In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, 783–794. ACM.

Christen, Peter. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Berlin: Springer.

"Digital Nautical Chart." 2020. 2020. https://dnc.nga.mil/.

Fellegi, Ivan P., and Alan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328): 1183–1210.

Feng, Chen-Chieh, and Alexandre Sorokine. 2014. “Comparing English, Mandarin, and Russian Hydrographic and Terrain Categories.” *International Journal of Geographical Information Science* 28 (6): 1294–1315. https://doi.org/10.1080/13658816.2013.831420.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Karimi, Hassan A. 2014. *Big Data: Techniques and Technologies in Geoinformatics*. CRC Press.

“Levenshtein Distance.” 2019. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=931218043.

Lynch, Maureen P., and Alan J. Saalfeld. 1985. “Conflation: Automated Map Compilation—a Video Game Approach.” In *Proceedings Auto-Carto*, 7:343–352.

ngageoint/hootenanny. (2015) 2020. C++. National Geospatial-Intelligence Agency. https://github.com/ngageoint/hootenanny.

Sarawagi, Sunita, and Anuradha Bhamidipaty. 2002. “Interactive Deduplication Using Active Learning.” In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269–278. ACM.

Sarawagi, Sunita, Anuradha Bhamidipaty, Alok Kirpal, and Chandra Mouli. 2002. “Alias: An Active Learning Led Interactive Deduplication System.” In *VLDB’02: Proceedings of the 28th International Conference on Very Large Databases*, 1103–1106. Elsevier.

Sun, Kai, Yunqiang Zhu, and Jia Song. 2019. “Progress and Challenges on Entity Alignment of Geographic Knowledge Bases.” *ISPRS International Journal of Geo-Information* 8 (2): 77. https://doi.org/10.3390/ijgi8020077.

Winkler, William E. 2002. “Methods for Record Linkage and Bayesian Networks.” Technical report, Statistical Research Division, US Census Bureau.

Acknowledgement:

Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).