On finding cross-lingual article pairs

Dirk Ahlers
Search Consultant

ABSTRACT

Finding a Wikipedia article in another language is often achievable with the in-built interlanguage links. We explore the possibility to automatically generate these links for geotagged articles as an application of entity resolution on an article level. It has the potential to improve Wikipedia, but also allows to use a well-curated ground truth for the merging algorithm. The resolution is based on only the simple features of coordinates and title. This is metadata that can be taken from APIs without parsing the full article itself. We use a conflation approach to identify articles with mismatched coordinates and a translation matrix tailored to the titles. Even complicated cases such as cities, municipalities, or departments with similar names at the same coordinates can mostly be identified correctly. Honduras was chosen as a test region because the country has a limited coverage (754 articles in both languages at time of writing [2]) that allows for a full manual assessment of results and because the resulting data is a basis for a geospatial search engine [1]. This finding has not been published in such brevity before, appropriate to the selection of features.

BODY

Cross-lingual merging of Honduran geotagged Wikipedia articles based on article names and locations alone results in 99.4% correct pairs.

REFERENCES

[1] D. Ahlers. Towards Geospatial Search for Honduras. In *Proceedings of the Latinamerican Conference on Networked and Electronic Media LACNEM 2011*, San José, Costa Rica, 2011. Universidad Latina Costa Rica.

[2] D. Ahlers. *Of 754 Wikipedia articles geotagged in Honduras, 345 are from the Spanish version, 409 are in English.*, 15.Jun.12, 7:52pm. Tweet. https://twitter.com/dirkahlers/statuses/213690505630990339.

Volume 1 of Tiny Transactions on Computer Science
This content is released under the Creative Commons Attribution-NonCommercial ShareAlike License. Permission to make digital or hard copies of all or part of this work is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
CC BY-NC-SA 3.0: http://creativecommons.org/licenses/by-nc-sa/3.0/.