Estimation of the Locations of the Language-Versions of Wikipedia - a Case Study on Geographic Data Mining (Abstract Version)

In this paper, language areas are estimated based on data of Wikipedia. About 10% of the articles in Wikipedia have a coordinate, i.e. a footprint. It is assumed that there is a relation between the languages spoken by the authors of an article and the footprint of the article. We assume that the author write about things that are near to them. This assumption is supported by the first law of Geography by Tobler: "Near things are more related than distant things".

To be more precise, the density of the footprints of a certain Wikipedia (that means a certain language version) is estimated. The regions where the density is high are well represented in that Wikipedia.

Yet, there are some objects which are very famous, e.g. *Arc de Triomphe* or *Eiffel Tower*, and because of this they are represented in a lot of Wikipedias. Further, there is a project in which the articles of Wikipedia are translated to other Wikipedias. Certainly, it is wrong to assume that so many languages (including Latin or Old English) are spoken at these places. Thus, the objects and their coordinates need a weight according to their famousness or probability to be translated.

In the articles of Wikipedia there are so called *Interwikilinks*. These links point to the corresponding articles in other Wikipedias. In the paper it was assumed that a object is famous or the article is probably translated if there are a lot of Interwikilinks in the Wikipedia article. Thus, the weight was defined as the multiplicative inverse of the number of Interwikilinks.

The result of this process is an image. In this image you can see, where the probability is high that a certain language is spoken and where it is low. In the figure you see an example for Pennsylvanian Dutch. You may recognize a hotspot in the East of the United States of America (namely where Pennsylvania is). This does not really surprise. But there are also some hotspots in the North of Spain or in the area around Luxembourg. This is not what may be expected.

Figure: Estimation of regions where Pennsylvanian Dutch is spoken (grey). The coastlines are included for a general orientation.

The estimations of the language areas have to be evaluated. In the paper the estimation of seven different Wikipedias using German languages: Written German, Alemannic, Bavarian, Ripuarian, Luxembourgish, Limburgian, and Dutch is presented. These estimations are compared to the language areas shown in language maps. A major problem of this comparison was to find a useful language map. It is shown that the estimations are quite reasonable.