

# Locating discussion board users with Bayesian analysis of geographic terms, language and timestamps



## Introduction

Online discussion boards have become an increasingly important communication channel over the last couple of years. These online discussion boards with people from all over the world discussing *en masse*, has become a new marketing channel. It is no longer up to the companies themselves to choose where and how to front their brands. The consumers are now doing this themselves, whether the companies like it or not. Suddenly, it is not just a few of your friends who are listening to what you have to say about your new phone, car or washing machine, but it may be thousands of people listening.

For companies and researchers working with analyzing this online mass communication and mapping users' behaviours and opinions, a method for finding the geographic location of the users would be a great asset. This thesis presents an effective method for doing this.

## Methods

This study is based on well established methods within the field of Pattern Classification and statistical analysis. The method uses Bayesian analysis as it's classification method.

Pattern classification is the task of placing different patterns into groups – in the context of this thesis; placing users into geographic locations.

## Proposed approach

The geographic location of discussion board users are in most cases not publicly available through the discussion boards. Because of this, other properties of the postings made by the

users have been considered. Three such properties has been chosen and a probabilistic approach has been applied in order to develop a method for finding the users' geographic locations.

**Time of posting (timestamp)** – The timestamp tells when the post was written. This can be used to find the most likely timezone from which the posting was made. By mapping geographic locations to timezones we can use this to estimate a probability of the posting being from the different regions. Figure 1 demonstrates how the classification of a user changes as the number of postings increases.

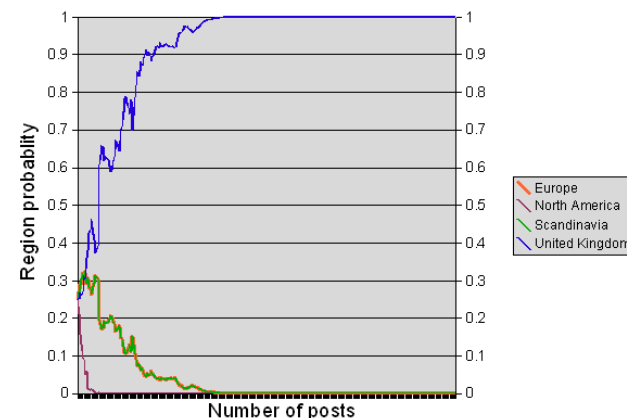


Figure 1: Classification of a user based on timestamp

**Language** – The language can give a good indication of the users' geographic locations, especially for non-English speaking users. By estimating the chance of users from different geographical regions using different languages, we can use this to find the probability of a user being from each of the regions.

**Mentioning of geographic terms** – Using the occurrence of geographic terms (names of places) as an indicator of geographic location

was chosen because users mention locations which may indicate where they are from. An example is “here in Norway the price is...”, which could be found in a discussion about cell phone prices around the world. This indicator has however been left out in the final prototype implementation due to poor performance and heavy resource consumption.

## Results

Figure 2 shows how the two properties, timestamp and language are used to determine a geographic location. This map shows the defined regions, which timezones they are located in and languages spoken in the regions. The regions are: North America (United States), United Kingdom (England, Scotland, Ireland), Europe (Germany, France, Spain) and Scandinavia (Norway and Denmark).

The figure clearly demonstrates how using the timestamp as an indicator can easily help distinguish between some regions, like North America and United Kingdom because these are located in different timezones. However, the method is unable to distinguish Scandinavia and Europe because these are located within the same timezone.

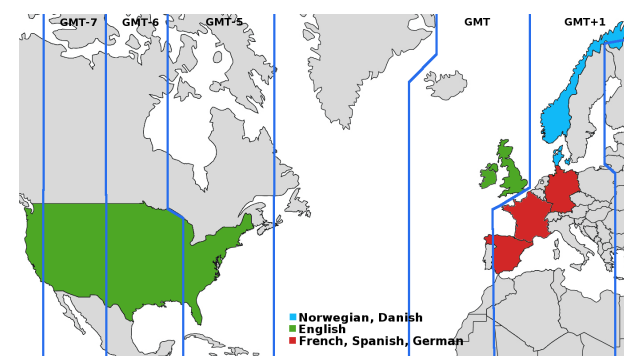


Figure 2: Mapping users to geographic locations

This is where the use of language identification shows it's strength. The analysis of the timestamp told that the region was either Scandinavia or Europe. By analyzing the language of the postings we can now find the correct region based on our knowledge about which languages are spoken in the different regions. This also works the other way around: The analysis of language is unable to distinguish United Kingdom from North America, but the timestamp analysis will be able to tell us which is the correct region, even though both are English speaking.

The prototype developed for this project has shown an accuracy of 88% - meaning that 88% of the users in the test sets were correctly classified.

## Conclusion

The results revealed by the prototype in this project has shown that the proposed method is working well and that the initial assumptions were correct: It is possible to determine the geographic location of users based solely on the information provided through their postings.

The proposed method has shown to be efficient and accurate, providing a new powerful tool for the field of analyzing the behaviour and opinions of users of online discussion boards.