Detecting Culture in Coordinates: Cultural Areas in Social Media

Christoph Carl Kling
University of Koblenz-Landau
56070 Koblenz, Germany
ckling@uni-koblenz.de

Thomas Gottron
University of Koblenz-Landau
56070 Koblenz, Germany
gottron@uni-koblenz.de

ABSTRACT

In this paper, we address the task of identifying the cultural relatedness of countries by extracting country specific behavioural patterns from social media. In a case study, publicly shared pictures annotated with spatial information are utilized to extract characteristic travel behaviour of different people in order to find areas of similar customs.

Categories and Subject Descriptors

H.2.8 [**Data Management**]: Database Applications—*Data Mining*

General Terms

Human Factors

Keywords

Culture, Social Media

1. INTRODUCTION

In Information Retrieval as well as Document Mining tasks, cultural knowledge is beneficial for determining relevance or for extracting useful information. Relevance may depend on the cultural background of the user seeking information and also of the author having written a document. In tasks such as sentiment analysis there is a large variety of how different cultures express their sentiments and opinions in a more direct or more subtle way. In a typical setting knowledge about the cultural background is hardly ever given explicitely, but only implicitely by document or user features.

One obvious piece of knowledge about culture is the language of a written text. But even when knowing the language of a document, the cultural context still plays an important role in identifying topics on a semantic level. Interpreting the opinion of a person referring to "Labour" is only possible when incorporating the knowledge that in several countries "Labour" is the name of a political party. Knowing the political borders of countries then is necessary to decide which of the Labour parties was actually referred to.

DETECT'11, October 28, 2011, Glasgow, Scotland, UK.

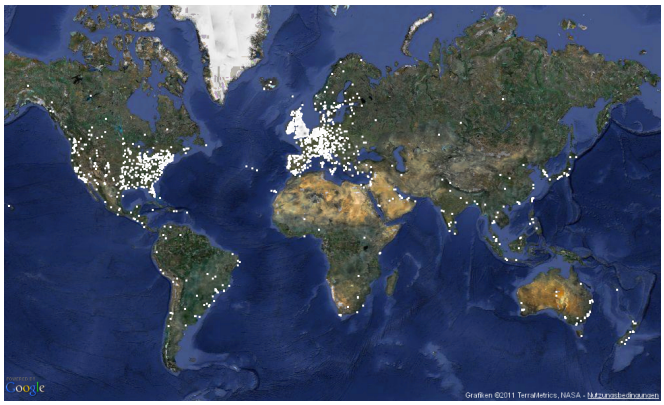

Figure 1: Geographic distribution of observed users

It is clear that cultural knowledge is crucial in many areas. The first step in identifying culture is to find areas of similar habits. These areas then can be compared to gain insights into the culture differences.

While there exist methods to detect cultural attributes such as language *based on implicit existing knowledge (e.g. characteristic n-grams [2]*) or based on questionnaires [3], to the best of our knowledge, there is no established method for an *automatic* and large scale detection of cultural areas.

In this paper we demonstrate a way of identifying cultural areas using geographically distributed data from social media websites and take a first step into the identification of cultural areas. We base our notion of cultural areas on the definition of a common behaviour of the members of cultural group. In a case study we exploit geografic data in social media data to mine intrinsic travelling behaviour of user groups.

The rest of the paper is organized as follows: in section 2 we provide definitions for cultural areas and look at related work in determining the boundaries of such areas or identifying common culture. In 3 we describe the composition of the data set and in 4 the feature generation and clustering methods we used for our analysis. We discuss the results of the clustering in 5 and conclude the paper in 6.

2. CULTURAL AREAS

The Oxford Dictionary defines culture as, among others: "the ideas, customs, and social behaviour of a particular

people or society". For our studies, we use this definition to gain insights in human cultures: whilst the ideas of a people can not be accessed directly, many customs and social behaviours are visible to researchers.

Views on culture which allow a numerical analysis emerged with the observation of cultural evolution. Karl Popper described the evolutionary nature of science, in which hypotheses are selected by falsification and favoured by their "truthlikeness" [8]. Science is only one part of culture, and Dawkins generalized the evolutionary view by introducing "memes", "a unit of cultural transmission, or a unit of imitation"[1], which he utilizes to construct culture as a result of "memetic" evolution analogous to the genetic evolution of species. Subsequent authors such as Lumsden and Wilson then introduced mathematical approaches to model the cultural evolution [5]. We take the evolutionary view on culture and extract traits in order to analyse the culture of different people.

Obviously, there exist areas of common culture between groups of people, which can be discovered in several ways. For instance, Hofstede used questionnaires from within a global corporation to extract cultural patterns across the earth [3]. Zelinsky utilised parts of company names to find cultural regions within a country [9]. Whilst the studies of Hofstede, to a large extent, are relying on personal interpretation, Zelinsky bases his study on neutral data. However, both authors rely on a relatively small sample size or on data of few information.

The availability of large amounts of user generated content carrying spatial information enables one to gain deep insights into behaviour of people at a large scale and with small effort [4]. Overell and Rueger analysed Wikipedia articles to extract regions of interest for different cultures identified by language [7, 6]. However, language based comparisons of cultures are limited because language is only one aspect of culture and there are many different cultures and subcultures sharing the same language.

In this paper we demonstrate how to discover patterns of similar behaviour across different people at any desired level of detail utilizing behavioural patterns observed in social media. In a pilot study we utilise annotated photos from a social media site to extract country-specific travelling profiles. Using these profiles we discover coherent cultural areas.

3. DATA COLLECTION

We chose flickr[1] as social media base to extract travelling behaviour. The advantages of flickr lies in its huge user base, media data annotated with geografic information and a good accessability via an API.

We bootstrapped our data collection with 4.169.974 geographically annotated pictures from flickr[2] being located within the area of Great Britain and Ireland starting from year 2000. These pictures were uploaded by 38.483 distinct users who at least once visited Great Britain or Ireland though not necessarily living within that area. By crawling their complete photo sets we obtained 51.098.373 photos, of which 19.319.106 had an assigned geo-location.

In order to derive travelling profiles for countries we evaluated the different user profiles. For 17.936 users we were able to assign a geographical position and country using the Google reverse geocoding API[3] on the given residence. We depicted the resulting locations on a map (see figure 1). The users originate from all parts of the world with a clear bias towards Ireland and Great Britain and countries whose citizens often visit the latter ones. This bias can clearly be attributed to our bootstrapping approach and can be handled to a large extent by normalizing the user distribution and discarding underrepresented countries.

We then determined the country in which the photos were taken by using the photo's geo-coordinates and assigning the country of the closest city from the Geonames[4] database. To calculate the distance between two geographical points $P1$ (e.g. the coordinates the picture was taken at) and $P2$ (e.g. the coordinates of a city), we used the *haversin* formula:

Let ϕ_1 be the latitude of $P1$, ϕ_2 the latitude of $P2$, $\Delta\phi$ the difference of both latitudes and $\Delta\lambda$ the difference between the longitudes of $P1$ and $P2$. Then the distance d of two points d on a perfect sphere with radius R is defined by:

$$d = 2 \cdot R \cdot arcsin\left(\sqrt{haversin\left(\frac{d}{R}\right)}\right) \tag{1}$$

where

$$haversin\left(\frac{d}{R}\right) = haversin\left(\Delta\phi\right) + \cos\left(\phi_1\right) cos\left(\phi_2\right) haversin\left(\Delta\lambda\right) \tag{2}$$

with

$$haversin\left(\theta\right) = \frac{versin(\theta)}{2} = \sin^2\left(\frac{\theta}{2}\right) \tag{3}$$

When comparing distances, the radius of the sphere is a constant and of no importance. Thus, we set R to 1.

4. DATA MODELLING

We obtained travel profiles for individual countries by creating a vector $c_i \in C$ for each country having $|C|$ dimensions where each dimension c_{ix} is the ratio of users from that particular country who visited country x. Since our sample set of users was created by crawling photos from Great Britain and Ireland, the geographical distribution of users in our dataset is biased. This oversampling of certain nations is counterbalanced by the normalization of considering only ratios and not aboslute numbers of users. However, also under-sampling poses a bias in the data. Whilst there are 11.214 users living in Great Britain, only 66 users are from New Zealand. We decided to limit our dataset to countries with at least 20 users in order to avoid bias by too sparse vectors. As a result we obtain a vector representation of the travelling behaviour for several countries.

Our hypothesis is that people of similar culture show a similar behaviour. Thus, it should be feasible to detect groups of culturally similar countries by clustering the vector representation. We use standard hierarchical clustering techniques and analyse the resulting tree of joined clusters. It is not necessarily useful to define static clusters of cul-

[1] http://www.flickr.com/
[2] http://www.flickr.com/
[3] code.google.com/apis/maps/documentation/geocoding/
[4] http://www.geonames.com

turally similar countries since cultural areas often overlap, show flowing transitions or contain subcultures.

Country vectors are compared using cosine similarity, rating two countries as similar when the distribution of visited countries is similar regardless of how frequent the citizens are travelling:

$$cosim\,(c_i, c_j) = \frac{c_i \cdot c_j}{\|c_i\| \cdot \|c_j\|} \tag{4}$$

We tried three different ways for agglomerative clustering:

- Single link clustering
- Complete link clustering
- Average link clustering

In single link clustering, two clusters Cl_1 and Cl_2 are merged together if the smallest distance between any of the elements of Cl_1 and Cl_2 is minimal. For culture areas this means that two culture areas are joined if they contain two countries which are culturally close. Since cultural borders are expected to show flowing transitions, clusters are likely to grow constantly even in directions of quite different cultures.

Complete link clustering fixes this problem by joining clusters where the maximum distance between the elements in the clusters is minimal. This implies that those cultural areas are preferably joined together, where every country of the one area is similar to all the countries of the other area. This way of clustering would perform best if there exist clear cultural borders and very homogeneous cultures.

Finally, average link clustering merges clusters of minimum average distance between all cluster members, being a compromise to the previous methods. In our case, cultural areas are merged if their average profiles are similar, so that typical attributes of cultures are determining the clustering process while single variations are of less importance.

Before starting the clustering process, we checked if similarities in travel behaviour are mainly caused by geographical conditions. To test the relatedness, we correlated the average distance between larger cities of at least 100.000 inhabitants of the studied countries with their cosine similarity of their feature vectors, using Spearmans Rho.

We found that there is a relatively small, significant ($p < 0.001$) negative Spearman correlation of -0.258 between geographical distance and our measure for cultural similarity. This shows that geographical proximity explains only a small percentage of travel behaviour.

5. CLUSTERING EVALUATION

We evaluated the clustering results, finding that single link clustering fails to find meaningful groups of similar countries, whilst complete link and average clustering show good results:

In figures 2 and 3 the clustering process is visualised by a tree whose leafs are countries. Numbers in the nodes show the order of clustering, meaning that higher numbers indicate less similarity of the joined clusters.

Figure 1 shows the complete link clustering process. First, Belgium and Netherlands are merged together, which are known to share a common history and which are – together with Luxembourg, which is not present in our data – widely perceived as a cultural area called "Benelux". Germany and

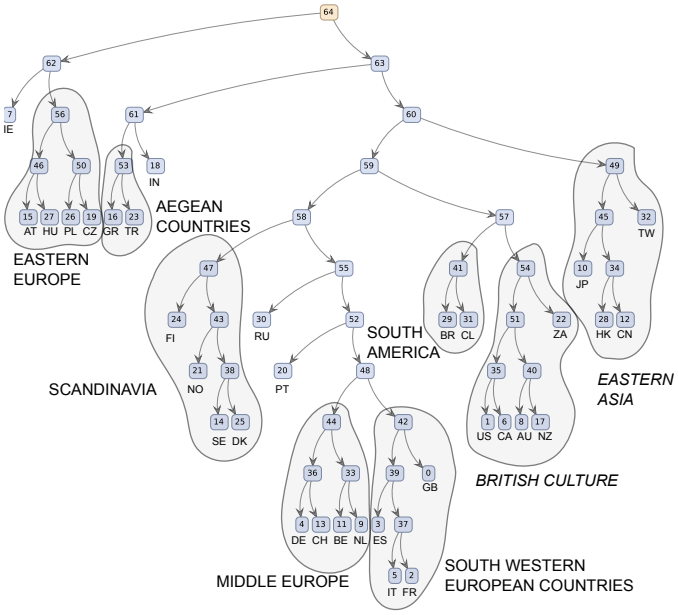

Figure 2: Hierarchical clustering of countries using complete link.

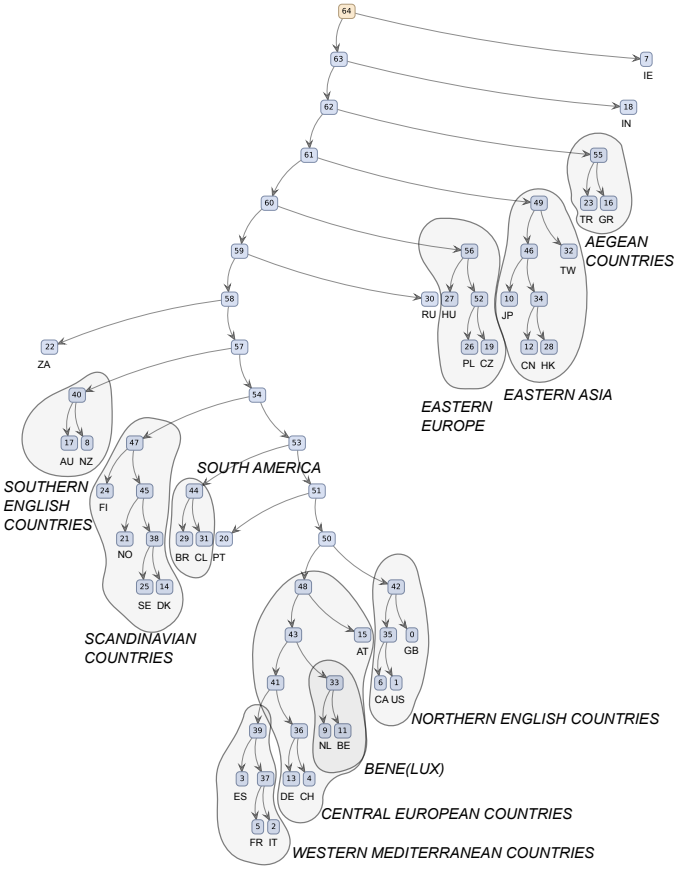

Figure 3: Hierarchical clustering of countries using average link.

Switzerland are merged, partly sharing the same language. They later get joined with Bene(lux), which then can be interpreted as "central european countries". The next joined countries are the Public Republic of China and Hongkong, which by now both belong to the PRC. Thailand and the Republic of China are added later, forming the cultural area of eastern asia. The next joined countries are the United States of America, Canada, Australia, New Zealand and South Africa, all together historically strongly influenced by being part of the british empire. India however is forming an own cultural area and Great Britain is joined with France, Spain and Italy. This is suboptimal, since subjectively the cultures of these countries seem quite different. A perfect cluster is the one of Finland, Norway, Sweden and Denmark, all together well known to form the Scandinavian culture. Hungary, Poland, the Czech Republic and most interesting Austria are joined together to form some sort of Eastern European culture. This is indicating that clusters are neither totally determined by geography nor language. Finally, South America and both countries at the Aegean Sea are joined. Notably, the Republic of Ireland is not joined with Great Britain or other countries, probably because of being a less popular travel destination and by that reducing the cosine similarity with all countries not visiting Ireland frequently.

When looking at figure 3 we see a slightly different clustering done by average link. Now Canada and the US are merged with Great Britain, whilst Australia and New Zealand form an own cluster and South Africa and India form own cultural areas. Also, Austria now is joined to Central Europe and Spain, France and Italy are separated from Great Britain. Scandinavia, South America, Eastern Europe and Asia are recognized again. This result subjectively seems slightly better than the previous one by complete link, since coherent languages are better grouped together. On the other hand, aspects such as the former affiliation to the British Empire are not recognized. This is quite natural, because cultural areas are changing their shape when focussing on different cultural aspects. Ireland again is not joined with any country till the late clustering phase in which all clusters are merged.

Overell visualized the frequency of referred geographical areas for wikipedia articles in different languages by displaying them on a world map [7] and interpreted this distribution as the geographical focus of different cultures identified by language. With our method we are able to display geographical distributions of culture attributes more precise, in our scenario for single countries. Figure 4 to 9 visualise the similarity of countries based on travel behaviour. Light green color shows high similarity, whilst red indicates least similar countries. Using these maps, one can discover interesting details, such as India being the least similar country for all European countries except for Great Britain and Portugal, which both have a history of colonisation in India.

6. CONCLUSION

We demonstrated how content from social media can be used to exhibit patterns of cultural areas. Though the used samples were heavily biased, findings already seemed mainly plausible. Using less biased data from different areas of behaviour should improve results significantly. Taking arbitrary vectors of behaviour extracted from social media which is assigned to a location enables us to find regions of similar

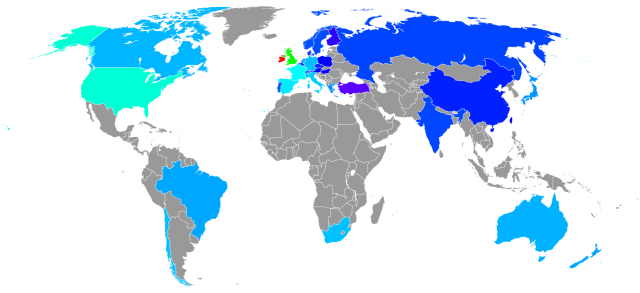

Figure 4: Cosine similarity between the vector of Great Britain and the vectors of other countries. Green indicates a high, red a low similarity. (This and following maps were created using gunn.co.nz/map)

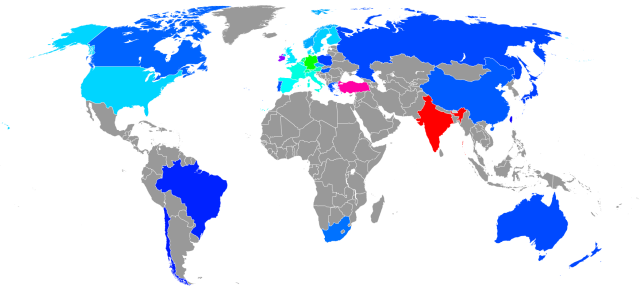

Figure 5: Germany

habits. If those regions are overlapping in many behavioural aspects, those regions can be interpreted as culture regions. Rich information sources such as texts or images then have the potential to exhibit which behavioural patterns are different across cultures in studies to come.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257859, ROBUST and grant agreement no. 248512, WeGov.

8. REFERENCES

[1] R. Dawkins. *The selfish gene*. Oxford University Press, 2006.

[2] T. Gottron and N. Lipka. A comparison of language identification approaches on short, query-style texts. In *ECIR '10: Proceedings of the 32nd European Conference on Information Retrieval*, pages 611–614, 2010.

[3] G. Hofstede. Cultural constraints in management theories. 1993.

[4] C. C. Kling, S. Sizov, and S. Staab. Virtual field research: A pilot case of biometeorology. 2011.

[5] C. J. Lumsden and E. O. Wilson. *Genes, mind, and culture: The coevolutionary process*. Harvard University Press, Cambridge, MA, 1981.

[6] S. Overell and S. Rüger. View of the world according to wikipedia: are we all little steinbergs? *International Journal of Computational Science*, 2011. In preparation.

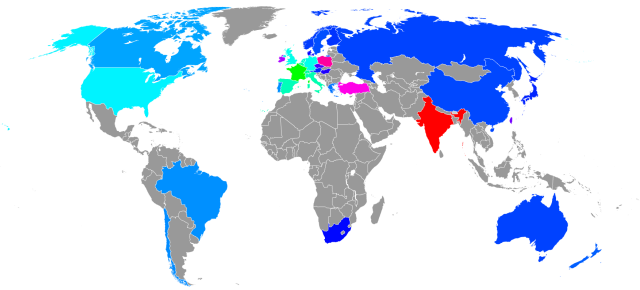

Figure 6: France

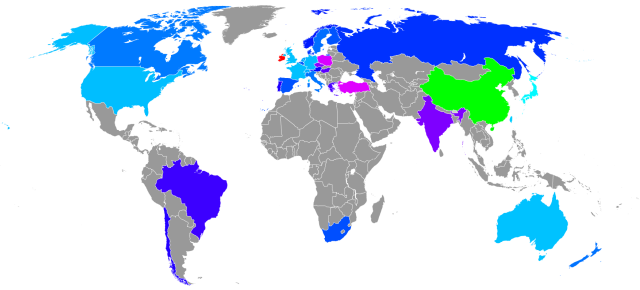

Figure 7: Public Republic of China

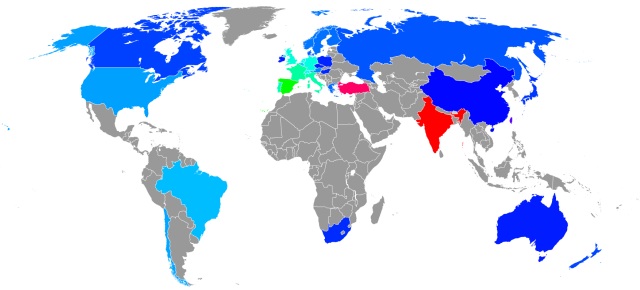

Figure 8: Spain

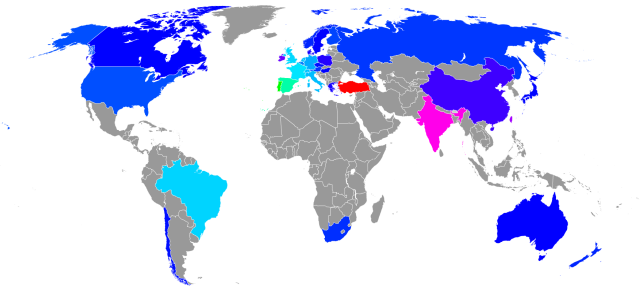

Figure 9: Portugal

[7] S. E. Overell. Geographic information retrieval:, 2009.

[8] K. R. Popper. *Objective Knowledge: An Evolutionary Approach*. Clarendon Press, Oxford, revised edition, 1979.

[9] W. Zelinsky. North America's Vernacular Regions. *Annals of the Association of American Geographers*, 70(1):1–16, 1980.