On the Quest of Discovering Cultural Trails in Social Media

Ruth Garcia
Web Research Group
Universitat Pompeu Fabra
Yahoo! Research Barcelona
Barcelona, Spain
ruth.garcia@upf.edu

ABSTRACT

With the constant increasing reach of the Web and in particular of Social Media, people create and share content that harbors information about habits, norms, preferences and values. Consequently, studying how *culture* influences users in online social media has increased the interest of several sectors such as the advertising industry, search engines and corporations. As a consequence, anthropological and computational models need to interact and complement each other to better target these new demands. Recently, several studies have analyzed culture from large-scale data but not many took into consideration the cultural models proposed by anthropological theory. By carrying out several experiments on large-scale data from the Web, we propose to combine theoretical concepts of culture with information technology techniques to process, analyze, model and interpret data from the Web. We plan to discover synergies between traditional social studies of culture and those derived from our experiments.

Categories and Subject Descriptors

H.2.8 [**Database Application**]: Data Mining; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation, Measurement

Keywords

Cross-cultural Differences, Social Media, Social Networks, Sentiment Analysis

1. INTRODUCTION

Defining the term *culture* is not a simple task and it can lead to several disagreements even for people who deeply study it. In fact, the most accepted cultural models lack homogeneity. Furthermore, the way culture is defined, studied and shaped can also vary according to its context, purpose or

function. A popular definition of culture comes from Hofstede: "culture is the collective programming of the mind that distinguishes the member of one group or category of people from others" [11, 13].

Due to the globalization and emergence of the Web 2.0, measuring and interpreting *culture* is imperative in several areas such as global information management, the internet advertising industry and web business applications [1, 26] as well as search engines that are attempting to provide more personalized search results. In fact, the now massive User Generated Content (UGC) in microblogs, social networks and other online spaces has given space to a huge repository of great cultural value. Although cultural information in UGC is inherently noisy, it can be extracted automatically and is plentiful. Moreover, the modern large scale analytic algorithms and tools available make its analysis and processing possible.

For all these reasons, the study of culture can be beneficial in several aspects. First, it can help identify synergies between cultural social studies manually generated and those automatically generated from "big data" on the Web. Second, understanding how culture affects user behavior online can give insights to user interface design and business strategy like e-commerce applications. In fact, previous studies showed that underestimating the impact of culture in user behavior has led some multinational corporations to fail when attempting to enter the market of new countries [23]. Third, modeling differences and similarities between cultures and how they change or remain the same in time can lead to the discovery of unexpected data (serendipity).

This PhD proposal focuses on exploring, analyzing and interpreting the role of culture in large scale UGC. Our methodology is to use Hofstede's theoretical basis for describing cultural characteristics. In order to extract, process and analyze features, we will use techniques from *data mining*, *sentiment analysis* and *network analysis*. Finally, quantitative user studies are also planned to be executed on groups of people from different backgrounds and cultures to observe and measure how culture affects online behavior.

2. BACKGROUND

2.1 Classification of Culture

One of the most influential works on modeling and defining culture comes from Dr. Geert Hofstede [10, 13]. He focused on how countries can be compared and contrasted in terms of only few values. He defended that culture is

a collective phenomenon but with varying individuals distributed as in a bell curve: "the variation between cultures is the shift of the bell curve when one moves from one society to the other"[11]. The values to compare cultures are called *cultural dimensions*. These dimensions were proposed as definition and they have a meaning *only* when making comparisons. Up to now, Hofstede built a cultural dimension model (CDM) made of six categories [11]. These dimensions are based on a broad empirical survey made in more than 70 countries. Relative scores are assigned to compare countries based on these dimensions and these scores only have a meaning if compared to other countries[1]. These dimensions are related to 1) the level of acceptance of unequal distribution of power, 2) individualism and collectivism, 3) masculinity and femininity, 4) the stress involved when dealing with uncertainty, 5) importance of tradition and 6) indulgence vs. restraint. These dimensions are seen more deeply in section 4.

In contrast to Hofstede's classification of culture at the national level, in the famous" The Clash of Civilizations?" [14], Huntington classified culture in a broader sense. He grouped cultures as civilizations, that is, the "highest cultural grouping of people and the broader level of cultural identity people have". He made an interesting point in identifying eight major civilizations based on their cultural differences: Western, Confucian, Japanese, Islamic, Hindu, Slavic-Orthodox, Latin American and possibly African civilizations.

2.2 Applications of Hofstede's work

There is not a clear and unified method to associate empirical data in different scenarios with Hofstede's dimensional cultures. Several methods have been adapted to specific situations. For example, Lipi et al. [18] measured how culture affects non-verbal expressions in conversations. They proposed a parameter based model employing Bayesian network techniques where culture is connected to Hofstede dimensions which are also connected with nonverbal expressions. For each node in the Bayesian network, probabilities are assigned according to Hofstede's CDM and to results obtained from experiments. Basically, when a country is chosen, the model outputs the estimated probability of behavior expressive parameters (Rigidness, Mirroring, etc). This approach forces the results to be probabilistically dependent of CDM.

Oh et al. [20, 19] also presented a model that integrates CDM and empirical data based on phone applications. First, they collected information about mobile-phone applications downloaded in several countries and classified each application according to its content (Finance, Entertainment, etc). Second, they defined a group of dimensions inspired in Hofstede and Schwartz as a basic framework for analysis. Then they used a Delphi Survey with 5-point Likert scales to assign values to each application according to the framework established by them. Next, for each phone application they calculated a *Cultural Dimensions Score of Content* (CSC) [19] based on the Delphi survey as to assign a CSC value to each application. Finally, they used this CSC score to calculate a *Cultural Index Score for Country*(CSIC) [20] which represents the country's inclination to a given cultural dimension.

Several techniques are used to present similarities and

differences of CISC values at the national level. Probably due to limited access to free downloaded phone applications, their results showed different characteristics from what Hofstede's and Schwartz have reported.

Other studies provided theoretical insights such as Chau et al. [4] who argued that individualism and collectivism are particularly relevant in studying the use of services built around Web 2.0, including OSN. Likewise, Ford et al. [7] discussed how to accommodate one of the five cultural dimensions in user interfaces to increase usability.

2.3 Cultural studies in OSN

Several studies reporting culture analysis of large-scale data from the Web have not followed a formal framework of cultural classification; however, the cultural insights reported have been interesting. For example, Poblete et al. [21] and Garcia et al. [8] carried out a study on understanding differences between the 10 most active countries in Twitter in the way they use social media, express emotions and interact with others. They identified correlations between language dependent and independent features. Yanai et al. [28], followed several algorithms to detect representative pictures from Flickr per region in the world. Furthermore, they categorized pictures and compare how photos of the same topic vary according to the country. Kling et al. [15] identified cultural relatedness of countries by assigning travel profiles to users in Flickr. They compared travel patterns of countries by identifying the home country of users and the countries where users have traveled. Hochman et al. [9] used around 550,000 images of Instagram from NY and Tokyo. They offered a comparative visualization research that indicates differences in color preferences, production, and varied hue's intensities which all shape a unique, local, "Visual Rhythm." Wang et al. [27] made an interesting cross-cultural comparison of the work load of scientists in America, China and Germany. They analyzed the day and time technical papers are downloaded from a popular scientific publisher's website. They concluded that there are some interesting differences but the common ground is that scientists are workaholics.

Table 1 summarizes the acronyms that will be used in this proposal:

Acronym	Meaning
UGC	User Generated Content
OSN	Online Social Network
CDM	Cultural Dimension Model
CSC	Cultural Score of Content
CISC	Cultural Index Score for Country
ANEW	Affective Norms for English Words

Table 1: Acronyms used

2.4 Hofstede in the Social Web

Our work differs from existing cultural research in multiple aspects: a) We plan to propose a model to represent cultural dimensions from UGC, user's interactions and network structure of *Online Social Networks*. b) we plan to use different time granularities to analyze culture and c) the context of our research is strictly based on large scale data from the Web which differs significantly from many of the previous cultural studies. To the best of our knowledge, there are not enough studies that associate Hofstede's model and cultural aspects of large-scale data.

---

[1]Comparisons of countries based on cultural dimensions can be dynamically obtained from Hofstede web page http://geert-hofstede.com/.

3. RESEARCH QUESTIONS

We propose to mine, process and analyze massive data likely to contain cultural information. We are not interested in comparing "specific nations/societies" but more interested in the process of modeling data to discover cultural differences so that the same process can be applied to any society/nation. We are also interested in reporting serendipity, that is, the non expected information discovery from our studies. In this regard, during the PhD, we will address the following questions in order:

- **How to model data from online social websites as to best represent Hofstede cultural dimensions?**
- **Can we find synergies between cultural data from online social websites and cultural findings from Hofstede?**

4. METHODOLOGY

In this study, we use Hofstede's six cultural dimensions as our basic framework for analysis. Some of the datasets we plan to use during the PhD can be extracted from public APIs while others can be provided by the laboratory where this PhD will be pursued. *We focus only on datasets with users' geo-location information(discovered by ips, locations reported by users, etc)* and profile description. Tentative datasets to be explored are 1) Twitter public API and other microblogs such as Sina Weibo, 2) Yahoo search/e-mail logs, 3) Yahoo! Answers and 4) Flickr. Whether to use all the listed datasets or all of them is something to be decided during the PhD [2]. Moreover, we plan to use Hadoop and Pig Latin to analyze data and other tools or dictionaries for language and geo-location detection.

For the analysis of cultural dimensions in large-scale data, it is important first to classify data based on the type of the analysis intended to be done. We propose two main types:

1. Language dependent: features specially related to sentiment analysis. In order to target sentiment analysis in different languages we propose to use the Affective Norms for English Words [2] and its adaptation to different languages [16, 22]. These dictionaries provide scores to the psychological reaction to which humans have to a specific word, according to a scale which ranges from pleasant to unpleasant. On the other hand, the Linguistic Inquiry and Word Count (LIWC) program is also used to calculate the degree to which people use different categories of words across a wide array of texts, and determine the degree any text uses positive or negative emotions, self-references, causal words, and 70 other language dimensions [25].

2. Language independent: features related to social network structure analysis, analysis of activity and interactions among users such as counting the number of conversational posts (i.e in Twitter it is possible to address a particular user with the symbol @ before the user name @username) .

Furthermore, we also propose to classify data on a temporal basis: per day, week, month or year. Temporal analysis is specially interesting in event coverage(protests, elections, holidays, etc) and evolution of user behavior.

---

[2]Currently, the research done so far has been made on Twitter

After making a distinction of the type of data to be studied and the temporal constrains to consider, we are now ready to address two main problems:

1. How to associate human behavior manifested online to Hofstede's cultural dimensions? This question arises when cultural behavior is identified and available in data and there is lack of consensus to associate findings to a particular cultural dimension.

2. How large-scale online social data can be modeled as to represent one or more cultural dimensions? We target this question when there is the hypothesis that data related to a particular cultural dimension can be found in online social activity such as in the affect of text, interaction with friends, network structure, etc. We attempt to explain cultural findings based on cultural dimensions.

In order to associate data to CDM we propose to follow an adaptation of the work done by Oh et al. [20, 19] and mentioned in section 2.2) use Delphi survey to determine the association of each CDM to the cultural characteristic/feature to be considered and 2) calculate a cultural index score per country for each specific dimension. The model will be adapted to web content, user interactions and social network characteristics. The following equations define each score (refer to [20, 19] for more information) :

$$CSC = \frac{\left(\sum_{i=1}^{j} *n_i \right)}{N} \quad (1)$$

$$CISC = \sum_{a=1}^{z} A \left(CSC * \left(\frac{dm_a}{DM} - \left(\frac{\sum_{a=1}^{z} \frac{dm_a}{DM}}{N} \right) \right) \right) \quad (2)$$

where A is a constant, a is the dimension based on CDM, dm_a is the total number of users at the national level with dimension a, DM is total amount of users with all dimensions and N is the total amount of target countries for the analysis.

We believe that the Delphi method is well suited to target lack on consensus when mapping data to cultural dimensions. The Delphi method is an iterative process that collect and distill the judgments of experts using a series of questionnaires. Each subsequent questionnaire is developed based on the results of the previous questionnaire. The process stops when consensus is reached or when sufficient information has been exchanged. Delphi surveys can be applied to problems that benefit from the subjective judgments of individuals on a collective basis [24]. This is the most appropriate solution we have found to map cultural data to CDM.

On the other hand, to target the second problem of modeling data to CDM, we have to give a closer look to Hofstede's cultural dimensions. With each definition, we present an example of how large-scale data can be related with CDM. It is still to be decided the definite data to be used.

- Power Distance (PD)- high vs. low: the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally. In a country classified as small-power-distance country, people would tend to easily approach and contradict their bosses. On the contrary, people from high-power-distance countries

will unlikely approach and contradict their bosses directly. For example, several studies have been made on the popularity of users in social media such as Cha et al. [3], one could explore the level of communication between users with high, average and low popularity for certain countries and relate these findings with PD scores.

- Individualism (IDV)- high vs. low: the level of integration into a group. Individualist societies (high IDV) represent loose ties between individuals where everyone is expected to look after him or herself and his or her immediate family. Collective (low IDV) societies represent individuals integrated into strong, cohesive groups where they protect each other in exchange of loyalty. For example, we can explore this dimensions in Twitter by measuring the amount of conversation exchanged with others as well as comparing the *unfollow* [17] dynamics of users.

- Gender (Gn) -masculinity vs. femininity: the description of roles between genders in a society. Feminine represents societies where emotional gender roles overlap: both men and women are supposed to be modest, tender, and concerned with the quality of life.Masculinity represents societies where emotional gender roles are clearly distinct: men are supposed to be assertive, tough, and focused on material success, whereas women are supposed to be more modest, tender, and concerned with the quality of life. For example, this dimension could be explored in social media by studying the use of adjectives and superlatives between women and men per country.

- Uncertainty avoidance (UA)- high vs. low: the extent to which members of a society feel threatened or uncomfortable in novel, surprising or unknown situations.This dimension can be measured by nervous stress , anxiety and the need for written or unwritten rules [12]. For example, measuring words expressing anxiety in social media can be an alternative to explore this dimension.

- Time orientation (LTO)- short vs. long: in essence, short-term oriented societies stand for the fostering of virtues related to the past and present (tradition, saving one's face, fulfilling social obligations) while long-term societies are more concerned with virtues oriented toward future rewards (perseverance and thrift). For example, this dimension could be studied in social media by identifying affiliations of users to certain groups. Many social sites are not only made of users but also groups representing ideologies, beliefs, etc.

- Indulgence versus Restraint (IVR): Indulgence is related to societies allowing free gratification where there is a high correlation between happiness, life control and importance of leisure as personal value. On the other hand, restraint refer to societies that control more gratification by means of strict social norms. In social media, we can combine studies made on happiness in blogs and microblogs such as in the work of Dodds et al. [5, 6] with this dimension.

5. RESEARCH TO DATE

We want to divide the years of research into the following stages: a) the first year to the analysis phase, b) the second year to the distillation of data into Hofstede's dimensions of cultures, c) the third year into using models to represent large-scale data into the form of Hofstede's cultural dimensions using one or more metrics and 4) the fourth year to report synergies and to replicate models into different datasets. During the first stage of our proposal, two experiments have been carried out to explore culture in microblogs. These experiments inspired the search of finding a model of culture than can be use in large-scale data. A summary of our finding are described in section 5.1.

5.1 Culture on the Use of Social Media

We conducted a large scale analysis of the behavior of millions of users in Twitter to observe significant differences among countries and the way they use social media during one whole year. More details about this work can be found in Garcia et al. [8] and Poblete et al. [21]. We covered the following categories: a) level of activity and languages used in the 10 countries that tweet the most, b) temporal happiness levels of tweets in two different languages for all active countries, c) the content of tweets in terms of mentions, hashtags, urls and re-tweets and d) the network structure: reciprocity, ties and social network.

Language independent findings are shown in the two right images in Figure 1. We can appreciate for instance that Australia differs notably from Indonesia's network. Indonesia has lower density and a larger core component which increases path lengths between nodes.

On the other hand, language dependent analysis is shown in he two right graphs in Figure 1. The graphs show the weighted happiness level per month when analyzing English and Spanish tweets. Note that the Spanish sentiment analysis is based on the ANEW list adapted to Spanish language by Redondo et al. [22]. We plan to carry out future studies using the Portuguese adapted dictionary of ANEW by Kristensen et al. [16] as well. In order to compute the "weighted average happiness level", we used the algorithm proposed by Dodds et al. [6, 5]:

$$h_{avg}(T) = \frac{\sum_{i=1}^{N} h_{avg}(w_i) f_i}{\sum_{i=1}^{N} f_i} = \sum_{i=1}^{N} h_{avg}(w_i) p_i \qquad (3)$$

where T represents all of the tweets per country for a particular time period and for a specific language (Spanish or English). Moreover, f_i is the frequency of the ith of N distinct words for which there is an estimate of average happiness (i.e., from the ANEW list [2]).

On the other hand, we studied the content and frequency of tweets and observed that the ratio tweet/user in Indonesia is the highest if compared to other active countries. Moreover, Indonesia also mentions other users in almost 60% of their tweets (highest value of all active countries). These insights suggest the collective nature of Indonesia. More research is needed to make comparisons using Hofstede's CDM.

6. ACKNOWLEDGMENTS

7. REFERENCES

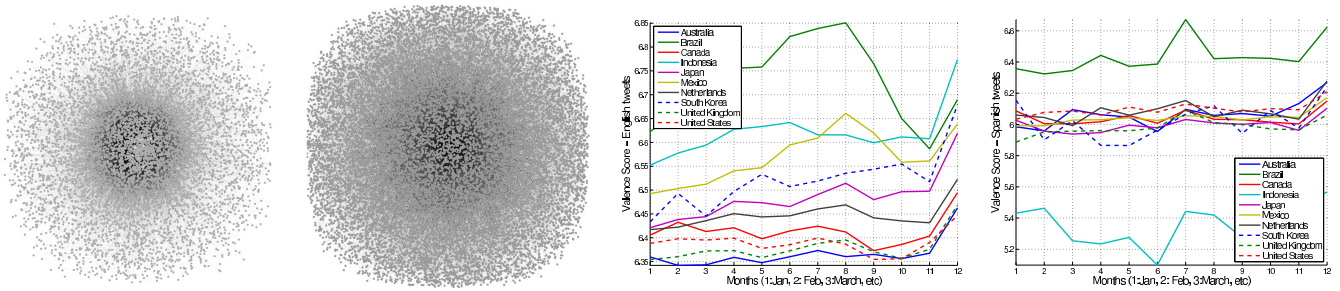[1] F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes. Social network analysis and mining for business

Figure 1: Network snapshot of Australia and Indonesia keeping edges of the same countries (left) and average happiness level for English Tweets per month for each country (right).

applications. *ACM Trans. Intell. Syst. Technol.*, 2:22:1–22:37, May 2011.

[2] M. M. Bradley and P. J. Lang. Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings. In *In Technical Report C-1, The Center for Research in Psychophysiology*, Gainesville, Florida, 1999.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *in ICWSM '10*, 2010.

[4] P. Y. Chau. Cultural differences in diffusion, adoption, and infusion of web 2.0. *Journal of Global Information Management*, 16(1):i–iii, 2008.

[5] P. Dodds and C. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, August 2010.

[6] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *Computing Research Repository abs/1101.5120v3[physics.soc-ph]*, Feb. 2011.

[7] G. Ford and P. Kotzé. Designing usable interfaces with cultural dimensions, 2005.

[8] R. Garcia, M. Medoza, B. Poblete, and A. Jaimes. Microblogging without borders: Differences and similarities. In *ACM WEBSCI'11, Koblenz-Germany*, June, 2011.

[9] N. Hochman and R. Schwarts. Visualizing instagram: Tracing cultural visual rhythms. Association for the Advancement of Artificial Intelligence, 2012.

[10] G. Hofstede. *Culture's Consequences, International Differences in Work-Related Values*. Sage Publications, Inc, CA, 1980.

[11] G. Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1), 2011.

[12] G. Hofstede, G. J. Hofstede, and G. Hofstede. *Cultures and Organizations: Software of the Mind*. McGraw-Hill, 2 edition, June 2004.

[13] G. H. Hofstede. *Culture's consequences : comparing values, behaviors, institutions, and organizations across nations*. Sage Publications, Thousand Oaks, California, 2nd ed edition, 2001.

[14] S. P. Huntington. The clash of civilizations. *Foreign Affairs*, 72:22, 1992.

[15] C. C. Kling and T. Gottron. Detecting culture in

coordinates: cultural areas in social media. DETECT '11, pages 11–16, New York, NY, USA, 2011. ACM.

[16] C. H. Kristensen, C. F. A. d. A. Gomes, A. R. Justo, and K. Vieira. Normas brasileiras para o Affective Norms for English Words. *Trends in Psychiatry and Psychotherapy*, 33:135 – 146, 00 2011.

[17] H. Kwak, H. Chun, and S. Moon. Fragile online relationship: a first look at unfollow dynamics in twitter. In *CHI 2011*, CHI '11, pages 1091–1100, New York, NY, USA, 2011. ACM.

[18] A. A. Lipi, Y. Nakano, and M. Rehm. A parameter-based model for generating culturally adaptive nonverbal behaviors in embodied conversational agents, 2009.

[19] J.-M. Oh and N. Moon. A cultural dimensions model based on smart phone applications. *JIPS*, pages 209–220, 2011.

[20] J.-M. Oh and N. Moon. Towards a cultural user interface generation principles. *Multimedia Tools and Applications*, pages 1–22, 2012.

[21] B. Poblete, R. Garcia, M. Mendoza, and A. Jaimes. Do All Birds Tweet the Same? Characterizing Twitter Around the World. In *CIKM 2011*, Glasgow,UK, October 2011.

[22] J. Redondo, I. Fraga, I. Padrn, and M. Comesaa. The spanish adaptation of anew (affective norms for english words). In *Volumne 39*, number 3, pages 600–605. Psychonomic Society Publications, 2007.

[23] D. Ricks. *Blunders in International Business*. Wiley, 2009.

[24] G. J. Skulmoski, F. T. Hartman, and J. Krahn. The delphi method for graduate research. *JITE*, pages 1–21, 2007.

[25] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. 2010.

[26] E. Vitkauskaite. Cultural adaptation issues in social networking sites. *Economics and Management: 2011.16 ISSN 1822-6515*, 2011.

[27] X. Wang, S. Xu, L. Peng, Z. Wang, C. Wang, C. Zhang, and X. Wang. Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics*, 6(4):655–660, 2012.

[28] K. Yanai, K. Yaegashi, and B. Qiu. Detecting cultural differences using consumer-generated geotagged photos. LOCWEB '09, pages 12:1–12:4, New York, NY, USA, 2009. ACM.