WORLD-IMPRESSION: How Do Netizens View the World?

Lei Zhang
Graduate School at Shenzhen
Tsinghua University
zhanglei@sz.tsinghua.edu.cn

Lizhi Wan
Graduate School at Shenzhen
Tsinghua University
lizhione@gmail.com

Thanassis Tiropanis
University of Southampton
tt2@ecs.soton.ac.uk

Wendy Hall
University of Southampton
wh@ecs.soton.ac.uk

ABSTRACT
In this paper we describe our ongoing project: the "WORLD-IMPRESSION" platform. The platform shows how people from a specific country view other countries by summarizing online data on the web. Opinions of web users viewing other countries are collected and updated regularly by crawling web forums and microblogs like Twitter. The database is then investigated using sentiment analysis and other data mining techniques. Visualization tools are used to distort the world-map proportionally to the attentions received by the countries. Tag-clouds are used to represent the top keywords describing a specific country.

Keywords
density-equalizing cartogram, tag-cloud

ACM Classification Keywords
J.4 [**Computer Applications**]: Social and Behavioral Sciences – *sociology*.

General Terms
Design, Experimentation, Measurement.

1. INTRODUCTION

You can see questions like "What do most Chinese think of Americans?" and "What do people in the United States think about Islam after 10 years of 9/11?" on Quora.com, a question-answering SNS platform. In general, it is important to know how people in one country think of other countries and even the whole world both culturally and politically [2][3]. Personally, people can have better understanding about themselves by looking at their impressions to others. Strategically, both the government and the society need to know this information in order to make more appropriate policies.

Figure 1. A joking example of how Americans view the world.

As shown in Figure 1, effective solutions to this problem require both statistical data and appropriate visualization tools for graphical interpretation of the results. Traditional ways like surveys and interviews have several limitations. First, it requires a large sample size for the data to be convincible, which is very cost-expensive; Second, data processing about the surveys are time-consuming considering the large number of sample size, which makes it very difficult for statistical work and data mining.

Therefore, we propose to use the Web as the data source for polling, processing and visualization of netizen opinions in this paper.

The benefits are as follows:

1. The Web has a much larger sample size then traditional surveys. As shown in Figure 2, netizens already represent a majority of the whole population in most developed countries. Even in less developed areas like Asia or African, the low Internet penetration rates don't necessarily mean a biased sampling of the whole population because of the fact that netizen age/income/occupation distributions resembles those of the whole population.

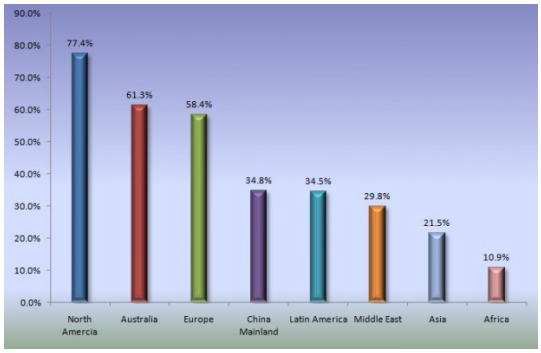

Figure 2. Internet penetration rates of different countries.

2. It's easy to crawl data from multiple countries and data sources automatically and regularly by employing multiple robots. Netizen opinions on different sources in different countries could be collected in a more efficient manner as well. Data could be stored and managed in large volume digital databases rather than on questionnaire hardcopies.
3. It's easier to perform sentiment analysis and opinion summarization. Digital database makes it possible to perform NLP (Natural Language Processing) tasks so that we could easily answer the questions we raised at the beginning of the paper.
4. Web visualization tools make it possible to distort the worldmap according to discussion frequency. By attaching top keywords netizens use to describe the countries, we can generate similar graphs like Figure 1.
5. It's easier to perform daily or monthly updates to reveal the dynamic changes of netizen opinions. It's also possible to observe the immediate response change of netizens when some very big events occur, 9/11 for example.

The rest of the paper is organized as follows. In Section 2, the framework of the WORLD-IMPRESSION platform is presented. In Section 3, we describe our current progress achieved, which is a simple prototype implemented with the data gathered for several major countries like US, UK and China. Some results and findings are also discussed in this section. Finally, Section 4 concludes the paper and discusses our future work plan.

2. PLATFORM FRAMEWORK

The main objectives of WORLD-IMPRESSION platform are as follows:

- **Comprehensive data**

  The system should be able to collect real and rich data from the Web. The data should be able to support query of opinions on how netizens in one given country view another given country. This requires data from all countries.

- **Effective analysis**

  The system should be able to provide effective analysis of data harvested from the Web. Sentiment analysis, for example, can be employed to decide the emotional attitude of the users and keyword abstraction methods could provide more details on the top words or phrases people use to describe a country.

- **Impressive visualization**

  A picture is worth of a thousand words. Therefore, appropriate visualizations are of great help to the presentation of the analysis results.

- **Full user engagement**

  A user-friendly interface should be provided so that UGC (user-generated-content) could be easily integrated into the system. For example, a user should be given the right to "agree" or "disagree" with the description of people's impression on a country by pressing "+1" or "-1" options provided by the system. A user can even add a new keyword if the system fails to harvest that keyword from the Web. This kind of feedback from the users could make the system to converge after a certain duration of time even if the description to a country might not be accurate at the very beginning.

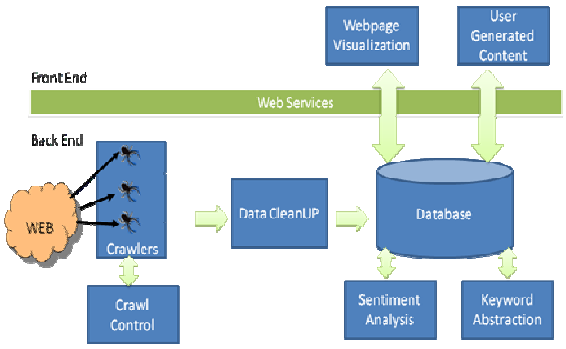

Figure 3. System architecture.

To achieve the objectives, we design the system architecture as shown in Figure 3, which includes the following major parts:

2.1. The Web

The Web is being used as a data source for polling netizen opinions. For each country to be studied, a list of websites should be carefully selected. We are specially focused on three kinds of websites: First, news media portals like BBC, New York Times, Xinhua News Agency, etc. which represent the mainstream authoritative opinions of one country to other countries; Second, social websites like Twitter, Facebook, weibo.com, etc. which represent the most up-to-date state of netizens in most countries; Third, web forums and BBS (Bulletin Board System), which represent university students in China. Data from these three different kinds of sources could be integrated to reveal the full picture. They could also be compared to one another by examining discrepancies among them.

2.2. The crawlers

Crawlers are robots or scripts which are sent to multiple data sources to automatically collect netizen opinions. These robots are intelligent so that they only crawl the pre-defined target websites for pre-defined target contents and won't overwhelm the whole Web. To enforce politeness, effectiveness, and fairness, crawlers are scheduled and controlled by certain policies.

2.3. Data Cleansing

When data have been crawled from the web, they may contain some unwanted contents, or have some flaws, or carry no meaning, which makes it impossible to reuse them. A cleansing procedure is needed to remove these low-quality data.

2.4. Data Analysis

Data analysis includes sentiment analysis, keyword abstraction and other effective data mining tools.

Sentiment analysis is needed for answering questions like "Do Chinese like the Great Britain?". By employing tools from the NLP (Natural Language Processing) field, sentiment analysis is able to determine the emotional attitude of the person by examining his tweets/posts.

Keyword abstractive is another useful technology in the system, which is able to abstract meaningful keywords from a long post/tweet. Frequency counts are recorded for the keyword dictionary so that we could display only the top keywords in the visualization step.

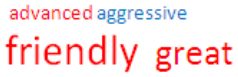

Figure 4. Opinions with sentiment, red/blue colors represent positive/negative feelings.

One challenge in the data analysis part is that we don't have enough expertise to process every single language in the world. Apparently, we cannot afford hiring Arab and Muslim agents and language experts as FBI does. Two feasible solutions could be: First, to collaborate with worldwide universities; Second, to use Google Translate as a proxy and perform analysis only after the target language has been translated to English.

2.5. Webpage Visualizations

There are many ways to visualize the data and findings, two promising methods of which are explained in details as follows:

- Map distortion

Figure 5 shows an example of map distortion according to global population where every country is enlarged/deflated according to their population proportionally. China and India become larger than they actually are because of their large population while North America and Europe shrink.

Figure 5. Worldmap distorted according to population.

In the WORLD-IMPRESSION platform, we employ the same idea to distort the worldmap with respect to the frequency of discussions of all countries mentioned in all target sources. The more people discuss a country, the larger it will be on the map. This gives the direct information of people's top concerns.

- Tag-clouds

In recent years, tag cloud has been widely adopted by web designers. A tag cloud is a visual representation

for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text (see Figure 6 for example). The font size of a keyword is proportional to its weight of significance. In this paper, the weight of a keyword is determined by its frequency counts in the data we harvest from the Web.

Figure 6. Tag-cloud example.

Finding the most appropriate way to display the tag clouds is crucial for user experience. One way is to open a pop-out dialogue box for tag-clouds when there's a mouse-click, which is easy to implement. Another way is to show them in the distorted worldmap, which requires more implementation effort.

By integrating map distortion and tag clouds, we will be able to generate visualizations similar to Figure 1, but with real and convincible data from the Web.

2.6. User Generated Contents

Although we try to be comprehensive, it is still impossible for the list of keywords to be exhaustive. Therefore, it important for the system to provide the users the opportunity to input new keywords if they feel the keywords are important enough. The users should be able to provide their feedback about existing keywords as well. Buttons of "+1" and "-1" are possible forms of incorporating user feedback. This feedback mechanism allows the system to converge to some common sense even if the description of one country is inappropriate at the beginning.

3. CURRENT PROGRESS

We implemented a prototype of the system, following the design goals and architecture we illustrated above.

3.1. Data Source

As a joint project between Tsinghua University and the University of Southampton, we are working on data from China and UK as the first step of the platform. In China, bulletin boards are very popular, especially in university students. They form a large social community on the web. For UK side, we recommend Twitter and Facebook as data sources. Due to the relatively heavy workload involved in Twitter API implementation, we currently only have some preliminary results about how Chinese netizens view the other parts of the world.

We crawled data from the following websites, tianya.cn and tieba.baidu.com. All the websites we study have their own search functions. We make use of them to crawl all the posts available without time-span restriction. Totally, we crawled more than 30,000 posts for 9 countries, as shown in Figure 7.

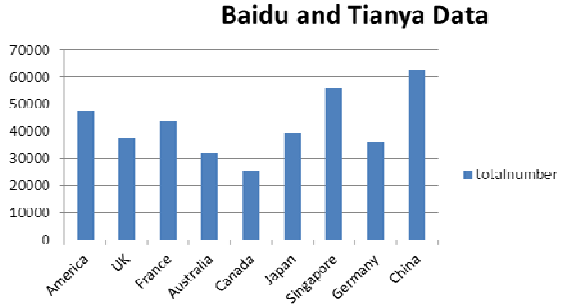

Figure 7. Current dataset size.

3.2. Data Cleanup

Before data analysis, we have to preprocess the data to make it cleaner. We have done the following procedures: extract sentences, filter the sentences according to some criteria. The output result is a set of sentences with other information such as time donating when comment is posted, where the source of the comment is and other related information. We use MySQL database to maintain the whole dataset.

Three interesting problems arising in the data cleansing step are as follows:

- Co-reference problem

When selecting the keywords, we pay special attention to the co-reference problem, i.e. one country may be referred to by different names. For example, "America" "USA" "US" "the States" all mean one country. We solve this problem by maintaining a table for country alias.

- Subset problem

People may not express their views on a specific country but the continent as a whole, like the Africa and Europe. In this case, one effective mention of a subset country will be considered an effective mention of the superset continent as well. However, one

effective mention of the superset continent will only be considered 1/N mention of each individual subset country, assuming the total number of countries in the continent is N.

- Default-referring problem

Some user might not mention the name of the country explicitly when replying to a thread. We assume that every post following the same thread is related to the same country mentioned in the first post of the thread.

3.3. Data Analysis

We try to employ the latest research accomplishment in NLP and data mining to get more accurate and meaningful results for our system. We divide data analysis into four subtasks: topic extraction, holder identification, claim extraction and sentiment analysis.

For topic extraction, there are simple cases. Take the sentence of "America is advanced." for example. "America" specifies the domain topic since it is explicitly stated in the sentence. For some other non-trivial cases, there are two strategies to attain the goal. One is use the feature of the phrases such as the noun position etc. The other is to use co-occurrence of the candidate topics and the context of the sentences.

For holder identification, we select the named entity strategy, which considers the user who posts the comment as the holder of the opinion. If the same user posts multiple comments in the same websites, the system will treat these comments as different and independent ones.

For claim extraction, we need to take gratuity into consideration. We assume every comment is related to a country and the comments somehow explain the holder's opinions. By using Chinese language segmentation tools and other features including semantic orientation, subjective words and phrases etc. the system extracts a rough opinion from the comments.

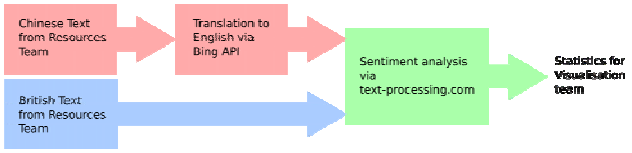

Figure 8. Sentiment analysis workflow.

With the help of sentiment analysis, the keywords and phrases are related to human emotions and attitudes. We use HowNet [4], a semantic library similar to WordNet, to calculate the tendency of the keywords in our platform.

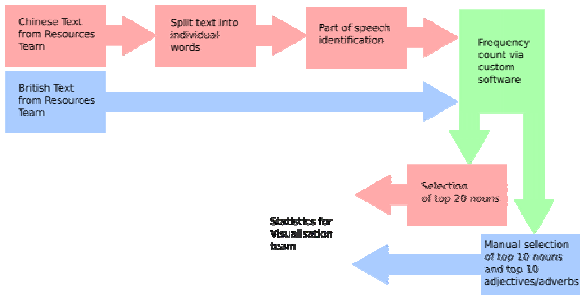

Figure 9. Keyword frequency count workflow.

3.4. Data Visualization

Data analysis outputs a pool of opinion tags, each associated with a frequency count. The data visualization step then organizes the top keyword tags into a tag-cloud. The color and font size of a tag are determined by the frequency counts and sentiment analysis results obtained in the data analysis step. Different color represents the sentiment tendency of the opinion. Font size represents the popularity of the tag. The whole map is distorted using MAPresso [1], according to the weights of the countries. The weight of a given country is obtained by summing the frequency counts of all keywords associated with the country. The result is shown in Figure 10. Here, we only show 9 countries whose territories are marked green.

Figure 10. Preliminary visualization results.

As shown in the figure, the world map is distorted due to different weights associated with the countries. It's obvious that UK becomes much larger than it actually is. We can observe similar things on Singapore.

For US, UK and Australia, tag-clouds are generated to show the detailed keywords that most Chinese use to describe them.

It's ironic that the first keyword that comes into Chinese mind when talking US is "Iraq". "Individualism" is the keyword with the second largest weight in the tag-cloud for US.

For UK, people in China seem to be very interested in the Royal Family, with "Prince Charles" "Diana" and "Kamila" in the top tags. "Darwin" is also often discussed because of his contribution in evolution theory.

For Australia, postgraduate education is of the most interest to most Chinese with almost every keyword in the tag-cloud related to it.

4. CONCLUSION AND DISCUSSIONS

In this paper, we briefly introduced our WORLD-IMPRESSIONS platform, which is an ongoing project at Tsinghua University and the University of Southampton. The project aims to use the large volume of data collected from the Web to facilitate the understanding of how people in different countries view each other. By employing data-mining skills and visualization tools, interesting results are demonstrated.

Future work will be conducted in the following aspects:

- Currently system can only deal with Chinese and English. Future work needs to deal with more languages.
- Twitter and Facebook APIs will be implemented to enrich our data sources.
- An *N***N* matrix will be built, each row and column representing a country. Element [*i*, *j*] contains the tag-cloud summarizing opinions of country #*i* to country #*j*.
- More user-friendly interface design will be implemented, including 3D tag-cloud, seamless integration of the tag-cloud and the map, zooming and highlight of the map, news/tweet display when clicking the mouse on a specific keyword.

ACKNOWLEDGEMENT

The authors would like to thank the fruitful work and discussions of the following students from University of Southampton during their visit to the Graduate School at Shenzhen, Tsinghua University: Paul Gaskell, Laura German, Richard Gomer, Christopher Hughes, Sarosh Khan, Terhi Nurmikko, Lisa Sugiura, Jiadi Yao, and Aristea Zafeiropoulou. We would also like to thank the following Tsinghua students for their contribution: Hao Chen, Long Cheng, Qi Li, Lizhi Wan, Jinchuan Wang, Tianxiang Yan, Bo Zhang, Mengfei Zhang, Jie Zhao, and Yiming Zhou.

REFERENCES

[1]. MAPresso. http://www.mapresso.com/.

[2]. Benjamin I. Page, Tao Xie, Living with the Dragon: How the American Public Views the Rise of China, June 2010, Columbia University Press, ISBN: 978-0-231-15208-2.

[3]. "iSpeak China": What are Young Chinese Thinking about?. http://www.chinasmack.com/2011/pictures/adrian-fisk-what-are-young-chinese-thinking-about.html.

[4]. HowNet, http://www.keenage.com.