How and why exploit tweet's location information?

Flavien Bouillot
LIRMM
16 rue Ada
Montpellier, France
flavien.bouillot@lirmm.fr

Pascal Poncelet
LIRMM
16 rue Ada
Montpellier, France
pascal.poncelet@lirmm.fr

Mathieu Roche
LIRMM
16 rue Ada
Montpellier, France
mathieu.roche@lirmm.fr

Abstract

Tweets exchanged over the Internet represent an important source of information, even if their characteristics make them difficult to analyze (e.g. a maximum of 140 characters, etc.). In addition associated to every message, lots of information such as location or date, are available. Taking into account these meta-information can be very useful for the decision maker. Obviously, due to the characteristics of tweets, relevant information are not expressed. For instance when considering tweets about natural disasters, an automatic extraction of location can significantly improve the analysis. In this paper, we propose a new approach to automatically extract and exploit the information about location. We also propose the end-user with a geographical hierarchy in order to improve the analysis of a set of tweets. Experiments carried out on real medical data underline the relevance of our proposal.

Keywords: Twitter, geographical hierarchy, disambiguation, tweet, multidimensional analysis.

1 Introduction

In recent years, the development of social and collaborative Web 2.0 has given users a more active role in collaborative networks. Blogs to share one's diary, RSS news to track the latest information on a specific topic, and tweets to publish one's actions, are now extremely widespread. Easy to create and manage, these tools are used by Internet users, businesses or other organizations to distribute information about themselves. This data creates unexpected applications in terms of decision-making. By default the data is placed in the public domain and thus can be used by anyone. Indeed, decision makers can use these large volumes of information as new resources to automatically extract useful knowledge.

Since its introduction in 2006, the Twitter[1] website has developed to such an extent that it is currently ranked as the 10th most visited site in the world[2] Twitter is a platform of micro blogging. This means that it is a system for sharing information where users can either follow other users who post short messages or be followed themselves.

[1] http://twitter.com
[2] http://www.alexa.com/siteinfo/twitter.com

In January 2010, the number of exchanged tweets reached 1.2 billion and more than 40 million tweets are exchanged per day.

Tweets are associated with meta-information that cannot be included in messages (e.g., date, location, etc.) or included in the message in the form of tags having a special meaning. Tweets can be represented in a multidimensional way by taking into account all this meta-information as well as associated temporal relations. In this paper, we consider the data warehouse [1] as a tool for the storage and analysis of multidimensional and historical data. Furthermore we focus on the standardization of location data to make them usable in a data warehouse. It thus becomes possible to manipulate a set of indicators (called measures) according to different dimensions that may be provided with one or more hierarchies. Associated operators (e.g., Roll-up, Drill-down, etc.) allow an intuitive navigation on different levels of the hierarchy.

In this paper, we define a way to exploit information of location provided with the tweets in order to navigate into a geographical hierarchy like *city* < *administrative divisions* < *country*.

For example, in Figure 1, we have the 3 hierarchies:

- *Los Angeles* < *California* < *United States*
- *Sacramento* < *California* < *United States*
- *Houston* < *Texas* < *United States*

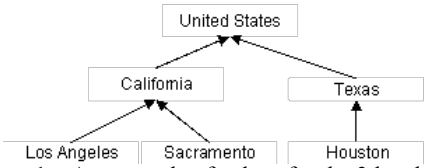

Figure 1 – An example of values for the 3 levels hierarchy in the United States

In tweets, information about location can be specified in very different ways:

- Set through an Internet access point (e.g., "London, Uk").
- Extracted from with geographical coordinates (latitude, longitude) when the tweet is sent by a mobile phone (e.g, "43.611, 3.877").
- Manually filled by the user. Here also lots of ways of expressing locations exist. Some of them are directly usable (e.g., "Paris, France" or "Usa"), or can be used after some text transformations (e.g., "L.A." for Los Angeles, "NYC" or "NY" or "New York" for New York City) but also it exists very useless information such as: "worldwide", "in Justin Bieber's bed" or "near a goat".

Finally, sometimes there are some tweets without any information about location and they will not be considered in this paper.

If we can not determine location (no match or no information), there is another way to determine the location by using the user's time zone. This indication is set automatically by Twitter but can be changed manually by the user according to a closed list of choices. In Twitter, the time zone is represented by a city (usually a capital city such as: "Paris", "Lima", "Quito" or "Rome").

Thus a French user will be attached to the "Paris" time zone while an Italian one will be attached to the "Rome" time zone although the time difference relative to the Greenwich meridian remains the same for France and Italy. Even if this information is not very precise it gives at least the country where the tweet was sent.

In this paper, our objective is to automatically extract this information, when available, to fill a hierarchy and then associate a tweet to a specific location in order to help a multi-dimensional analysis.

The rest of the paper is organized as follows. Section 2 describes our approach to extract and normalize the location from the all the information provided in tweets. In Section 3, we present some results of conducted experiments in the medical domain. Before concluding by presenting future work in Section 5, we discuss about these results in Section 4.

2 Method

As we said in the introduction, location information can be either text or geographic coordinates. In this section we introduce how to deal with these heterogeneous information.

2.1 Baseline

From text, we first have to determine if the provided information is useful or not. Furthermore, if it is possible one problem remains on the automatic assignment of tweets on the hierarchy (e.g., "*Paris*" is a city and "*Usa*" is a country).

With geographic coordinates, the most interesting because the most accurate, it requires a different approach because two people outside of 100 meters will not have the same coordinates.

To enrich the information about locations, in conducted experiments, we choose *Geonames*[3] as a reference. *Geonames* is a geographic database which is accessible free online under a Creative Commons license and contain over 8 million geographical names corresponding to more than 6.5 million of existing sites. These names are classified into 9 categories and 645 sub-categories. Data such as latitude, longitude, elevation, population, administrative subdivision, zip code are also available in several languages for each location.

We have incorporated cities with a population of over 1000 people to limit the volume of our baseline. Moreover, in the case of homonyms (e.g., *Paris < Ile de France < France* or *Paris < Illinois < United States*), we have decided to keep the cities with the largest population[4].

We have thus decided to keep 88,574 cities in which we added some common abbreviations ("*Washington*" for "*Washington DC"*, "*JAX*" for "*Jacksonville*", "*OKC"* in Oklahoma City, etc.).

Furthermore five aliases referencing five different time zones available in tweets have been integrated with the coordinates of the major cities according to their area: ("*Central Time (US & Canada)*" is attached to the city of "*Chicago"*, "*Eastern Time (US & Canada)"* to the city of "*New York"*, "*Mountain Time (US & Canada*)" to the city of "*Tucson"*, "*Pacific Time (US & Canada)*" to the city of "*Los Angeles*" and "*Atlantic Time (US & Canada'"* to the city of "*Fredericton"*).

At the end we have a database references the cities with 88,586 elements.

Two other databases are used. The first-one references the country and contains the mapping between names and common abbreviated name (e.g.; "*usa*" and "*us*" for "*United States of America*", "*uk*" for "*United-Kingdom*") and the mapping between the name in English and local language name (e.g.; "*Spain*" and "*España*").

Since administrative divisions differ between countries. The second-one references these divisions. For example we consider the state for United-States and Australia, the province in Canada or the home-nations for the United-Kingdom.

This architecture with 3 databases (City, Administrative Divisions and Country) allows us to precisely determine location from tweets:

- Geographic coordinates
- Manual input of a city name, state, or country
- Time zone

Furthermore it would be easy to add another level of our hierarchy by using another database containing eg street names, county or province.

[3] http://www.geonames.org/

[4] We discuss of the disambiguation issue in Section 4.

2.2 Process

Now, we present the main process defined to extract location. First of all we first analyze the content of the tweets to extract relevant terms that could correspond to some location specification. This step is performed by using specific patterns specifically defined. Then we consider the location information from the meta-data. We then try to identify the geographic location from the location information and, if such an information is not provided we consider the time zone. When geographical coordinates are available, we compute a distance between the geographical coordinates of the tweet and the geographical coordinates of the city from the following equation proposed in [2]:

$$R \times \operatorname{acos}\left(\cos(LaT) \times \cos(LaV) \times \cos(LoV - LoT) + (\sin(LaT) \times \sin(LaV))\right)$$

With:

- R = 6366 (stands for the radius of the Earth in km)
- LaT = latitude of the Tweet in radians
- LaV = latitude of the City in radians
- LoT = longitude of the Tweet in radians
- LoV = longitude of the City in radians

We hold the city for which the distance between its coordinates and those of the tweet is the lowest.

Finally when coordinates are not provided, we extract any word preceding commas. First words are usually sufficient to find the location (e.g.; "Los Angeles, United States"). Then we query the database to find some cities that can be candidate. If one of them exists we affect this value at the city level of the hierarchy and get the information from the database to affect the other values city, administrative division, and country we found in the city database to the tweet. Otherwise we query the division database to find one potential area and then affect the values administrative division and country extracted from the database to the tweet. In such a case a wild character (symbolised by *) is affected to the city. Finally, if all the previous operations fail a query to the country database is performed to affect country value to the tweet. Insofar, a * is affected both to the city and administrative levels of the hierarchy. In some tweets it is possible that all the previous information are not available, we thus focus on the information provided by the time zone. In that case, we apply the same process by requesting the relevant database and affect the corresponding value to the country level while the * is affected to the two lower levels.

The main drawback of using this last issue is that we can only affect a tweet to one country. We agree that this affectation could appear less relevant but according to the analysis that can be performed we have notice that the results obtained remain quite relevant. This issue is addressed in the following section.

3 Application to Medical Domain

In order to evaluate our approach of extracting location, several experiments were conducted. They were performed using PostgreSQL 8.4 and some Perl scripts. We illustrate our proposal by querying in real time Twitter with medical terms in a tweet stream thanks to the Twitter's Stream Api[5].

In order to focus only on medical terms, as medical reference, we have chosen the *MeSH* [6] (Medical Subject Headings) taxonomy which is used for indexing PubMed articles. The Stream Api imposes a limit of two hundreds keywords. To extract the tweets related to the vocabulary used in *MeSH*, we focused on the tweets related to the "Virus Disease" (MeSH ID: C.C02) and queried Twitter by using all the terms of the corresponding sub-hierarchy.

This sub-hierarchy consists in 363 words including 198 unique terms among which we find generic terms such as "*virus*" or "*influenza*" and more specialized terms such as "*HIV-associated lipodystrophy syndrome*" or "*feline Acquired Immunodeficiency Syndrome*". Due to the limitations of the 140 characters on tweets, a term such as "*feline Acquired Immunodeficiency Syndrome* " does not have any chance to be written in a tweet. We thus decided to focus on the 198 terms reduced to only one word in the hierarchy. Finally, tweets were treated as in [3].

Extracting the location, as explained in the previous section, allows us:

- To locate on the map the number of occurrences of specific diseases. For example we can observe in Figure 2, the distribution of the use of the *leukemia* word in all the tweets over the world (in this figure we do not consider tweets coming from US and Canada).

Figure 2 – An example of visualisation of the tweets distribution for a specific term

- To navigate within the hierarchy by using OLAP operators. For example, it is thus possible to know the set of the most specific terms for a city, a region and a country.

For our experiments, we have collected 2,495,122 tweets from the 21 January 2011 to the 20 May 2011. From these, 1,842,569 tweets were provided with location information (73,8%). Among this 73,8%, our analysis showed that the process to efficiently assign location for at least the country level is a success in 75% of cases. Of these 75% success rate, over 43% of locations can be made with the city or geographic coordinates as shown in the distribution in Figure 3.

[5] http://dev.twitter.com/.
[6] http://www.nlm.nih.gov/mesh/MBrowser.html

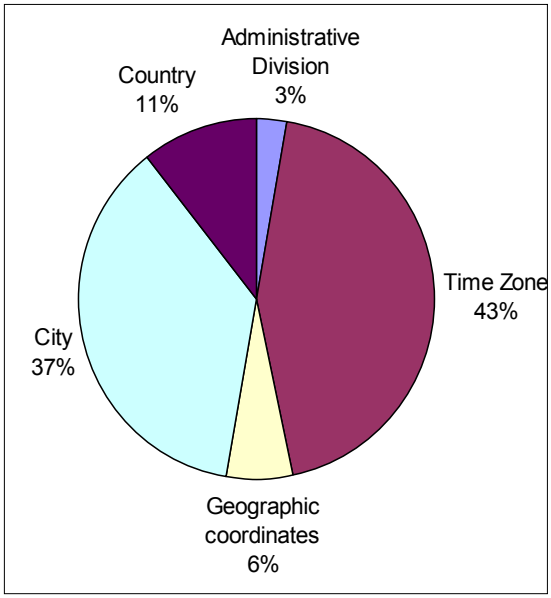

Figure 3 – Distribution of the source of the information used to determine the location.

4 How to disambiguate location

We conducted a number of choices that need to be discussed.

As we said in the introduction, we have decided to not consider city homonyms (e.g.; Paris, France and Paris, Illinois United States). Similarly homonyms can also exist in different levels. For example Mexico is both the name of the country and a city. In such a case, we decided to remove the reference to the city level and affect the tweets to the country level.

In our experiments,16.4% of the 1,842,569 tweets refer to a city with multiple geonames inputs. 2.513 different cities are covered in this volume of 302,609 tweets.

In Table 1, we illustrate the top 20 location homonyms in this huge volume of tweets.

location	Number of tweets	% of tweets	location	Number of tweets	% of tweets
London	16,259	0,88%	Boston	4,188	0,23%
Mexico	9,865	0,54%	San Diego	4,061	0,22%
Los Angeles	8,335	0,45%	Sydney	3,914	0,21%
Bandung	6,745	0,37%	Florida	3,488	0,19%
Toronto	6,255	0,34%	Houston	3,465	0,19%
Atlanta	6,152	0,33%	Buenos Aires	3,216	0,17%
California	5,872	0,32%	Santiago	3,213	0,17%
Washington	5,453	0,30%	Miami	3,135	0,17%
Manila	5,360	0,29%	Dallas	3,111	0,17%
San Francisco	4,707	0,26%	Melbourne	2,973	0,16%

Table 1 – Percents of homonyms

Some methods of city disambiguation is presented in [4]. An other solution for dealing with homonyms has been evaluated. This approach was based on the language using in tweets. To do so, we have used the TextCat software. TextCat is an implementation of a text categorisation algorithm based on the N-grams principle [5]. This technique consists in extracting all the N-grams (i.e. sequences of N consecutive characters) and then counting the occurrences of each N-gram in the text. For example, the 3-gram "*the*" occurring very often in the text is characteristic of the English language.

Basically, we can infer that a tweet having a lot of occurrences of French words is likely to be posted from Paris, the main city in France, than in Paris in Texas.

Nevertheless the main drawbacks of this approach is that:

- A recent study in 2010 done by *Semiocast* [7] has showed that 44% of tweets posted from France were written in French and 34% in English. This observation is also relevant for other places. For instance, since only 42% of posts in Italy in Italian (we can notice that 95% of tweets from Japan are in Japanese).
- Even if this approach is appealing, obviously when the same language is used in different countries, it becomes useless. For instance, in London, Minnesota, USA and in London, United Kingdom every tweets are expressed in English or American English.

One way to solve the last issue is to perform a natural language analysis of the content of the tweet in order to automatically detect some expressions used either in UK or in USA.

Unfortunately, note that with 140 characters, there are only on average of 8 words in a tweet. This problem and the technical constraints that have to be considered, i.e. performance and data volume, are too complicated to manage compared to the expected gain that we did not yet addressed this issue.

For solve this kind of problem we plan to apply text-mining techniques for identifying geographically-aligned lexical variation directly from raw text. For instance the work of [5] is based on a supervised approach (K-Nearest Neighbors) in order to predict the location. This one corresponds to the average of the positions of the K most similar authors by using a similarity measure (i.e. *cosine*) from a training set of tweet data ($K=20$ is chosen in the experiments).

5 Conclusion and Future Work

In this paper we proposed an approach to geographically locate tweets in order to use the OLAP operators used for a multi-dimensional analysis. We thus had focused on the different way to automatically extract from tweets the relevant information and gave some experimental findings. We also highlighted the problem of homonymy and proposed some solutions. To further improve our approach we want to extend the proposed approach to take into account both words used in the tweets but also the stream of tweets for each user. Thanks to this history we would like to improve the process by taking into account the information extracted from the last tweets sent by the same user. Furthermore by improving the analysis of the text we plan to extract relevant information that can be

[7] http://semiocast.com/downloads/Semiocast_500_000_tweets_par_jour_sont_emis_en_France_20100331.pdf

useful for inferring the location. For instance, if a tweet is about the Eiffel Tower, there are great chances that the tweet comes from Paris, France, rather than Paris, Texas [6].

In future work and in order to disambiguate homonyms location in the geographical hierarchy, we plan to adapt the $AcroDef_{MI}^{3}$ method described in [7]. This measure is based on the Cubic Mutual Information [8] that enhances the impact of frequent co-occurrences of two words in a given context.

In our case, we plan to calculate the dependence between a location to disambiguate and different words of the tweets using the context of the hierarchy (i.e., parents (Administrative Division, Country) of the location).

References

[1] Codd, E., Codd, S., Salley, C., Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. *White Paper,* 1993

[2] Vincety T., Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations, *Survey Review XXIII* (176): 88–93, 1975 Retrieved 2009-07-11.

[3] Bringay S, Béchet N, Bouillot F, Poncelet P, Roche M, Tesisseire M., Towards an On-Line Analysis of Tweets Processing, In proceedings of *DEXA 2011,* 2011.

[4] Paradesi S, Geotagging tweets using their content, *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, 2011.

[5] Cavnar, W. B. and J. M. Trenkle, N-Gram-Based Text Categorization, *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13 April 1994.

[6] Eisenstein J., O'Connor B., Smith N.A., Xing E.P., A Latent Variable Model for Geographic Lexical Variation. *EMNLP* 2010: 1277-1287, 2010

[7] Roche, M., Prince, V., Managing the acronym/expansion identification process for text-mining applications. *Int. J. of Software and Informatics* 2(2) 163-179, 2008

[8] Daille, B., Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *PhD thesis*, Université Paris 7, 1994