Extraction and Classification of Twitter Messages to Apply in Business Intelligence

De La Cruz S. Carlos and Cortez H. Paola

Abstract—Nowadays the participation of users in information networks is increasing. Every day millions of people share opinions, experiences, modes through a comment, etc. However, this source of information called collective intelligence is unutilized at its maximum capacity, because there comes a point where the user is oversaturated with so much information. The reason would be advisable to have the means to exploit the information that is expressed through the comments so that is synthesized and allow support to collective needs. This paper presents a way to exploit the information found within tweets of Twitter through programming language, i.e., applying data mining to the comments made. The aim is to show that through this language is simple start analyzing what the users say. The business intelligence system through social networks is proposed such as a case of study, which shows how they can be used in comments to support making decision. The Social Web has had a drive through the various websites that allow you to share information and which are known as social networking. The participation of users across these networks is astonishing; actually there are millions of people around the world who participates daily with at least one comment. But considering that a simple comment may contain relevant information, and joined the thousands of comments at a given time there is a source of collective information, synthesize that information would be important to support collective needs. This source of information can be manipulated in order to find patterns in the words that are used in the comments, and use them as benefits for the users.

Index Terms—Information networks, collective intelligence, Twitter, business intelligence.

I. INTRODUCTION

Nowadays the social Web had had an impulse across the different Web sites that allow to share information which is known as social networks. As a result, the participation of users across these networks is amazing, because millions of people in entire world every day participate at least with one comment.

However if we considerate a simple comment can express outstanding information, and joined the thousands of comments to the instant it has a collective information source; it could be important to synthesize this information to support collective needs. That source of information would be manipulated, with the purpose to find pattern in useful words in comments and use it to benefit to the users.

In Mexico the top ten in which is concentrate the biggest numbers of users in social networks are: Mexico State, Distrito Federal, Jalisco, Veracruz, Nuevo León, Puebla, Baja California, Guanajuato, Chihuahua and Tamaulipas. Have been found that Mexican people use mainly the social networks to communicate between buddies; to follow and give opinions about culture, entertainment and sports; know about the last news and meet other people [1]. In this moment the most popular social networks useful in the world are: Facebook, YouTube and Twitter. In this moment México is located in seventh place among the countries with more twits, we are talking about almost 15 million of users; the ranking is United States with 140 million of “cuentas” in the social network [2].

Undoubtedly knowing the reason which people use the social networks and importance to find ways of synthesize information, below is an instance how using Python and the API of Twitter may get information. The principal aim is take advantage of information that every day is sharing with other people, and it manages to help collective needs. The study case is a business intelligence system based in social networks, which it searches patterns of words with the propose to do a classification of Twitter comments.

What is happen? Everyday million of users answer this question through the most popular site Web called Twitter which is concentrated a huge amount information, as consequence of active participation of their users. It is because through 140 characters the user is able convey information which is propagated of exponential form. Only enough with write information about every event for this known in a few minutes by a lot of users. However, those little messages called Tweets, can find outstanding information that allow: exemplify the profile of user, characterize geographic areas, or just know the current topics of interest at a given time. Twitter is not define as a social network, instead as a information network in real time that connect users with the last stories, ideas, opinions and news about more interesting [3].

According to a study made by [4] México is located in the seventh place between the countries which have more tweets in the world because have almost 15 million of users, Having every day million of tweets generate by users. Therefore that users have with a lot of information which is not easy to assimilate. For this reason is necessary to employ techniques that may find information that be useful, like data mining.

Data mining is a process that have as a purpose to find out, extract and keep information about huge data bases, through
search programs and identification of patterns and global relations, tendencies, deviations and other indicators apparently chaotic that have an explication, it would find out across different techniques with this tool. [5] if Twitter is consider as a huge source of information, due to realizable to apply techniques of data mining to explorer information.

Moreover in [6] it mentioned that Twitter's API join with Ruby allow get useful information and more synthesized. In this paper we expose the way to obtain information using Python and Twitter's API, with the purpose to do a classification made by users.

II. FRAMEWORK

Especially in these days talking about interchange information in social networks had increased eventually, and this is the principal reason of the investigation. The comments that users post or share is interesting to know, because comments have a knowledge source called collective intelligence.

The increased or necessity to communicate by means of computer, it have allow more interchange of information. In [7] is called some advantages to use Twitter hence to have a limit of characters for writing an idea which is necessary be clearly and concise. The authors mentioned the fact that social networks participate people who had interest in common, also with different experience. One of the advantage to have information Web is let informal communication, as a consequence increase the possibilities of interchange information very valuable, make a confidence source between users.

In [8] it had mentioned that usually Web sites allow interchange of information and let analyze collective behavior, because is hard analyze user individual information. That way is important the fact that users need to be linked in social networks, and it happened because they have some interests or characteristics in common.

On the other hand when we analyze a comment is necessary be an aim, and in [9] is indicated that words are indentified and can be keys to establish a connection between the place where user had done a comment. This result very incredible because user feels more close with the next services that can be interested. In [10] also in comments can have similar words that help to identify in which have made the user this comment. In [11] is using an information network with the purpose to find a characteristics and behaviors of specific geographical area after classify, not paying attention to comment semantic.

Based on the research presented, it is important to mention that the analysis of key words about a comment may get certain semantic level. Accordingly in the moment that is identify the user idea, it can propose a different ways to auxiliary or analyze these information the right way.

In the next part, it mentioned the way is possible get information using API of Python and Twitter, afterward it had showed an study case with the aim to classify comments.

III. EXTRACTION THE INFORMATION

When the information is shared in social networks, this information is transmitted to user's friends or contacts, it is supposed that users have similar profit and likes. in this way the user information that he shares can be use as a result of others contacts. Fig. 1 is an example where shows the connection between Web friends.

Last figure had seen the relations inside a social network, when user makes a comment, these information is send all the list of contacts. Whence to noticed a propagation level that produce a comment, is interesting to know what is shared.

To check the information that is found such social networks as different Web sites, it is noticed clearly that this information is based in comments. At the moment to do a searching through various Web sites, the majority of the occasion is getting recommendations which are based in the parameter of popular vote disappointed the express opinion by the users. It will be doable analyze the semantic content of comments, with the target to understand what user wants to express and synthesize information than is found in social networks and various Web sites, to be more easy the interchange of information. This allows have a big get close called semantic Web.

A. Get Information with Python

Moreover to do this analyze the fist was investigate about the way get information since Twitter APPI, also we noticed that Python is a programming language which uses a couple of sentences that allow establish the communication with the Twitter count of user.

For these proof had used a Python Package Manager, because this language has characteristics such as be clearly in his syntax, it had Object Oriented Programming, had huge kind of data and is compatible with C/C++, .NET, and others.

For instance it is showed some principal instructions necessaries for getting information from a user count in Twitter. For this proof had taken a [13] as a reference.

When we write in a console of Python the next instructions, it establishes the communication with API of Twitter:

```python
>>> easy_install twitter
>>> import twitter

>>> twitter_api =
twitter.Twitter(domain="api.twitter.com",
    api_version='1')
```

If the connection is successful, it can prove when we need
to print the last comments of users contacts. but it has a little problem in these proof (this happened if is using Java), the problem is that only obtain the last 20 comments, and it is impossible have old comments.

It is possible to manage API of this information and know in this little moment about users are saying through trend topic; which is generated by user quantity that they use words in their comments. For instance, the next instructions show the results when is possible to get information about trend topic using at specific moment:

```python
print [trend['name'] for trend in trends[0]['trends']]
```

However the result of this information is more complete if we know it: who and where is talking from trend topic in special. Afterward to make a selection about a trend topic in specific, it gets users who has commented it. This information result very detail, it presents data as geographical coordinates where the comment was made also user name. To show the way how users appear, we get the next results:

```python
>>> freq_dist.keys()[:50]  #Las 50 palabras mas frecuentes
[u'RT', u'#MiHabilidadEs', u'que', u'la', u'en', u'a', u'y', u'de', u'el', u'no', u'me', u'las', u'para', u'@Jausensi:', u'fingir', u's\xe9', u'@HogwartsX:', u'como', u'horas', u'tu', u'@Twitter', u'@WhatsApp...y', u'donde', u'estar', u'estudiar', u'hasta', u'illega', u'nada', u'sin', u'ver', u'vez', u'...', u'@JulianSerrano07:', u'@Facebook', u'@mi', u'ai', u'cuando', u'd\xeada', u'unas', u'24', u'@BbTeDoe:', u'@DeboConfesarQue:', u'@DisfrutaLaVida4:', u'@TipicoNosotro', u'@WhatSappArg:', u'@WhatsApp Esp:', u'@cockhungryboy:', u'@dimeMister:', u'@FuckiuBITCH:]
```

To continue with this topic, it mentions study case for giving an example the use to do a manipulation adequate of information.

### IV. UNITS

Other manner to get information without any programming language is making a query directly since an Internet browser, because is possible make queries automatic way, the only think that is necessary to know is Id of a place cause break down the comments by people. To put this Web site address: http://twitter.com/search?q=place%3A247f43d441defc03, it has already made a query that result as an show in the Fig. 2.

However the result of this information is more complete if we know it: who and where is talking from trend topic in special. Afterward to make a selection about a trend topic in specific, it gets users who has commented it. This information result very detail, it presents data as geographical coordinates where the comment was made also user name. To show the way how users appear, we get the next results:

```python
>>> freq_dist.keys()[:50]  #Las 50 palabras mas frecuentes
[u'RT', u'#MiHabilidadEs', u'que', u'la', u'en', u'a', u'y', u'de', u'el', u'no', u'me', u'las', u'para', u'@Jausensi:', u'fingir', u's\xe9', u'@HogwartsX:', u'como', u'horas', u'tu', u'@Twitter', u'@WhatsApp...y', u'donde', u'estar', u'estudiar', u'hasta', u'illega', u'nada', u'sin', u'ver', u'vez', u'...', u'@JulianSerrano07:', u'@Facebook', u'@mi', u'ai', u'cuando', u'd\xeada', u'unas', u'24', u'@BbTeDoe:', u'@DeboConfesarQue:', u'@DisfrutaLaVida4:', u'@TipicoNosotro', u'@WhatSappArg:', u'@WhatsApp Esp:', u'@cockhungryboy:', u'@dimeMister:', u'@FuckiuBITCH:]
```

To continue with this topic, it mentions study case for giving an example the use to do a manipulation adequate of information.

### V. STUDY CASE: INTELLIGENCE BUSINESS SYSTEM IN SOCIAL NETWORK

The social networks is not more than evolution of ways to communicate between human being, than it has advanced with use of new channels or tools, it had based on collective knowledge across the users comments.

The companies for having publicity had been used social networks, and it had more popular nowadays is called media social, this process is doing publicity through social networks. Social media enable that clients are getting close to the
company, so the company be more close than clients and listen to them. Moreover thanks to this way to communicate in which is more easy to monitoring by user’s comments, the aim is to do automatic classification from comments making. This is the reason it is in process to development an invent than get a classification about the comments making a commercial mark.

For this case had been analyzed for two moths the comments making and marks as Coca-Cola, Nike y Revlon. Te target was determinate what follower share. As a consequence is important to mentioned that clients use this way to express the quality or give some recommendation about the product in which the comments were saying such catalogues as good or bad [14].

After this be a classification about comments , the next step is making a sub-classification of this, the result is show in Fig. 3. Using an ontology has achieved conceptualize the classification of comments.

A way to determinate the theory than is compose the ontology before mentioned, also is necessary to explorer API Twitter with the purpose to get words more usually used between by users. As a result the theory in Table I, which appear at the moment to do a comment positive to a commercial mark.

Currently in this study case it is found in the automatic classification phase of the comments. It is consist in determinate the quantity of classification of concept as "good" and "bad" that appear in user had written, in this way if appear more concepts negative inside the comment, it will separate as negative.

Finally when the comment has been classified, it produced to make a report such as: to indicate the quantity of comments good by time, to know if it expresses the users’ suggestions or complaints, etc. The goal is to companies generate automatic this reports in real time because they need what users are talking about it. It is enough to know that is talking about it, is necessary to know if is positive or negative.

VI. CONCLUSION

The information network has had an increase favorable with the active participation of users across the comments. Hence the result is a huge source of collective intelligence that should be explorer to get closer that it would be a semantic Web. The reason which this work is focused in show a way to get information using Python and Twitter API. For example is showed a study case where it uses this tools, to exemplify a business intelligent system based in social networks, which has to do a classification or Twitter comments. This is the first approach to make emphasis that the information express in comments should be using for exemplify the behavior of the participants, or provide indicators to support users, and thus assist in the collective needs of the people who participate daily in a information network.

ACKNOWLEDGMENT

The authors of this paper thank the National Institute Polytechnic (IPN), for the facilities provided.

REFERENCES


Carlos De La Cruz was born on January 23, 1974 in the state of Oaxaca, México. He has a master degree in Computer Science with specialization in Software Engineering from the National Center of Research and Technological Development located in the city of Cuernavaca, Morelos in Mexico.

Since 1998, he has been worked in Professional Interdisciplinary Unit in Engineering and Advanced Technologies (UPIITA) of National Polytechnic Institute (IPN), it is located in Avenue Instituto Politécnico Nacional No. 2580 Col. La Laguna Ticoman, Del. Gustavo A. Madero, CP. 07340 México,DF. His principal activity is as a research professor at the Academy of Telematic, who teaches and directs research and technological development projects. These investigations derived titling projects so that students can get the engineering degree.

Paola N. Cortez H. received the master in computer science from Computer Research Center of the National Polytechnic Institute (IPN). From 2007 to actually, she is research professor of Interdisciplinary Professional Unit of Engineering and Advanced Technologies (UPIITA), IPN. Her research interest includes the user profiling analysis and the social network analysis based on the users’ comments in order to find collective intelligence.