International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Mining Real Time Tweets for Cover Picture Suggestion by Text Mining Techniques

Nilam M. Zalavadiya[1]

[1]*Student, Dept. Of Computer Engineering, Darshan Institute of Engineering & Technology, Gujarat, India*

Abstract - *Web text (content) mining which is the part of web mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. Here web content is to be taken as Social network dataset of Twitter. The dissertation target is to generate word cloud that will be useful as cover picture for user profile of twitter. The term word cloud is a visual representation for text data, typically used to depict keyword metadata (tags) on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color. Word cloud is simply generated by past popular tweets of user profile. Intensity and importance of each tag is decided based on priority of popular tweets which is generated by sentimental analysis of every tweet. Novel algorithm is to be used to assign weight and color of tags based on popularity parameter of tweets. The proposed system will generate word cloud that provide through summary of all past popular tweets of user.*

Key Words: **text mining, twitter data mining, content mining, real tweets mining, keyword extraction, cover picture suggestion, web mining**

1. INTRODUCTION

Currently, online social networks such as Facebook, Twitter, Google+, LinkedIn, and Foursquare have become enormously popular all over the world and play a considerable role in people's daily lives. People access OSNs using both traditional desktop PCs and new rising mobile devices.

In recent years, online community/social networks (OSNs) have noticeably expanded in status around the world. In accordance with the data in October 2012, Facebook has 1.01 billion users using the site every month.

The rapid development of OSNs has involved a large number of researchers to discover and study this popular, ever-present, and large-scale service. In this paper, we focus on understanding user behavior in OSNs. OSN user behavior covers various social activities that users can do online, such as friendship creation, content publishing, and profile browsing, messaging, and commenting. Remarkably, these activities can be valid or malicious. So, understanding OSN user behavior is key to different Internet entities in numerous aspects:[1]

• For Internet service providers (ISPs), as OSN traffic is rising quickly and fetching significant, they want to learn the growth of the traffic pattern of OSNs. This can lead them to do several infrastructural actions (e.g., adding traffic optimization in network middle-boxes).[5]

• For OSN service providers, it helps them recognize their customers' attitudes to different functions, particularly for some tentative functions. Furthermore, from the perception of infrastructure investment, like which locations are cost-effectual to build data centers or else which content delivery network (CDN) cluster could be leveraged to distribute frequently accessed data, accepting users' geographic distribution as well as traffic activity is vital.[5]

• For OSN users, performance study is important to improve user experience. For instance, there are number of malicious accounts in OSNs. These accounts create unwanted messages for genuine users. So, detecting and blocking malicious users are important to guarantee good user experience.[5]

Today is an age of Social Media, where lots of websites are existing that are linking people worldwide, ex. Facebook , twitter, LinkedIn, etc. Among that Trend of Twitter analysis is becoming very popular nowadays. For example, AnalyzeWords helps disclose your personality by seeing at how you use words. It's based on good technical research connecting word use to which people are.

The service quickly gained worldwide fame, with more than 100 million users who in 2012 posted 340 million tweets per a day. The service also switched 1.6 billion search queries per a day. In 2013 Twitter was 1 of the ten most-visited websites, which has been known as "the SMS of the Internet". In May 2015, it has more than 500 million users, from which greater than 302 million are active users.

Research shows that people are using Facebook and Twitter for connecting with friends and brands. They're now looking to these platforms for updates on current events.[2]

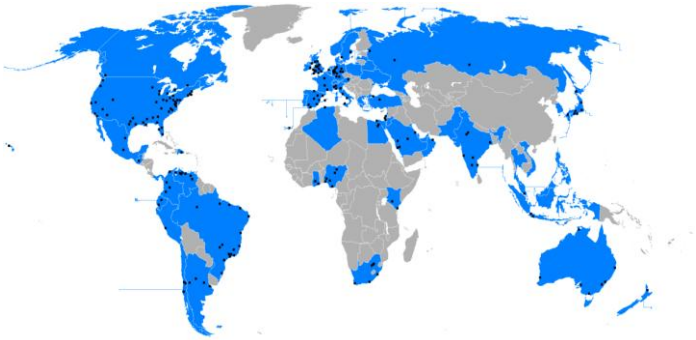


Fig -1: Twitter Popularity among Worldwide Countries[9]

Thus Considering Social Media – Twitter as part of analyzing user personality and also Considering the amount of information published in social networks, and the growing speed in which this information is updated and exchanged, our research focus is about how interactive information visualization techniques can help in the relation and extraction of – new – information from the original ones.

1.1 Web Content Mining & Its Emerging Trends

Web content mining is that the scanning and mining of text data, pictures and graphs of an internet page to find out the connection of the content to the search query. This scanning is completed once the clustering of the web pages through structure mining and make available the results based mostly upon the level of relevance to the recommended query.[6]

Emerging trends of web content mining are as follows:

- **Data/information extraction:** Our focus is on extraction of structured records from Web sites, like products and search results. Extracting such knowledge (data) allows one to grant services. Two main forms of techniques, machine learning and automatic extraction are enclosed.
- **Web information integration and schema matching:** Though the WWW contains a huge amount of information, each web site (or even page) presents the same information differently. How to determine or match semantically similar data may be a vital problem with several practical applications.
- **Opinion extraction from on-line sources:** There are several on-line opinion sources, e.g., client reviews of services & products, forums, blogs and chat rooms. Mining opinions (especially customer opinions) is of vast importance for marketing intelligence and products benchmarking.
- **Knowledge synthesis:** Concept hierarchies or metaphysics (ontology) are helpful in various applications. Howexver, generating them manually is extremely time-consuming. Some existing ways that explores the data redundancy of the WWW will be accessible. The key application is to synthesize and systematize the pieces of data on the Web to convey the user a coherent image of the subject domain.
- **Segmenting Web pages and detecting noise:** In various Web applications, one solely wants the most content of Web page with no advertisements, navigation links and copyright notices. Automatically segmenting the content of the Web pages to extract the major content of the pages is attention-grabbing problem.

1.2 Introduction to Word Cloud

With the advent and great success of a new generation of community-oriented websites in domains such as Media Sharing (e.g., Flickr, YouTube) or Social Bookmarking and Citation (e.g., Delicious, Connotea), a new way of metadata creation has emerged, commonly known as tagging. Tagging-based systems enable users to add tags – freely chosen keywords – to Web resources to categorize these resources for themselves and/or others. [10]

Fig -2: Word Cloud Example

Visual browsing within the tag collections is accomplished in numerous ways; usually, websites proffer an interface element called tag cloud. Generally, a tag cloud presents a definite range of the most frequently used tags in a defined area of user interface. A tag's popularity/frequency is expressed by its font size and is thus simply recognized. Sometimes, additional visual properties, like font colour, intensity, and weight, are manipulated. Next to their visualization function, tag clouds are also navigation interfaces because the tags are typically hyperlinks resulting in a collection of items they're related to.[11]

2. APPROACH OVERVIEW

The architecture of my system is schematically illustrated in Fig -3.

1. Dataset of Twitter is to be given as input to proposed system. Dataset mainly contains nodes with user tweets of social media that are significantly connected through Link.
2. To mine user tweets of nodes Text Mining algorithm is to be apply.

3. Algorithm of Text Mining will extract out popular words from tweets using technique of parsing.

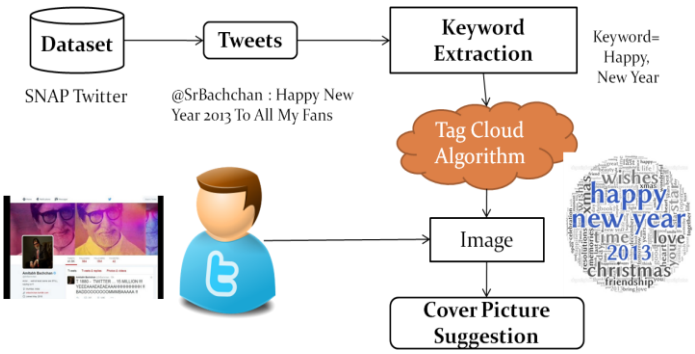


Fig -3: Proposed Architecture

4. Once Procedure of Mining of Post Of users is performed that will extract out popular keywords from post.
5. As and once keywords are fetched, next process is to perform Tag Cloud Generation Algorithm on that keywords.
6. Tag Cloud Generation Algorithm will create Image of those popular words.
7. At the end this Image will be recommended as Community to set as cover picture.

3. SYSTEM IMPLEMENTATION

3.1 Tools & Libraries

- **Import.io API:** Import.io is a web-based platform for extracting information from the websites with no need of writing any code. The tool allows individuals to create an API by using their point and click interface. Users navigate to a web page and teach the app to extract knowledge by highlighting examples of data from the page, learning algorithms then generalize from these examples to figure out a way to get all the data from that website. The data that users gather is stored on import.io's cloud servers and user can download that data as CSV, Excel, Google Sheets or JSON and shared. [4]

- **Rapid Miner- Analysis & Evaluation Process:** Rapid Miner is a software system platform developed by the company of the similar name that offers an integrated environment to machine learning, data processing, text mining, predictive and business analytics. It's used for business and viable applications plus for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process together with data preparation, results validation, visualization and optimization.

- **OpenNLP:** The Apache OpenNLP library is a machine learning based toolkit for the dealing out of natural language text. It supports the foremost common NLP tasks, like tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co reference resolution. These tasks are typically needed to create more advanced text processing services.

- **NetBeans:** NetBeans is a software development platform written in Java. The NetBeans permits applications to be developed from a collection of standard software components called modules. Applications that support the NetBeans Platform, with the NetBeans integrated development environment (IDE), can be extended by third party developers.[12]

- **JAVA:** Java is a general-purpose computer programming language which is concurrent, class-based, object-oriented, and particularly designed to encompass as few implementation dependencies as possible. It is meant to let application developers "write once, run anywhere" (WORA), which means that compiled Java code will run on all platforms that support Java while not the necessity for recompilation. Java applications are normally compiled to bytecode that can run on any Java virtual machine (JVM) despite the consequences of computer architecture. As of 2015, Java is one of the most accepted programming languages in use, particularly for client-server web applications, with a reported 9 million developers.[8]

- **OpenCloud Library[7]:** OpenCloud is a Java library for generating tag clouds, conjointly referred to as weighted list. The two main classes within the library are the Tag class that represents a single tag and the Cloud class that represents the word cloud in its entirety. The Cloud class act as a collection that you can populate by adding Tag objects.

  Each tag has a score value that shows its level of importance. Tags with a high score will be given a higher weight. When the tag is added to the Cloud object, if it is already present a tag with the same name, their scores are summed. Since the default score value is 1.0, if don't specify score values, total score of a tag equals to the number of times that it's been added to the Cloud (frequency of occurrence of the tag).

  The Cloud class converts the scores to weight values using a linear equation. A user can decide the range of values that the weight can have, so the weight values can be suitably used for tag cloud visualization.

3.2 Important Method of OpenCloud

To create tag cloud there is must requirement that Tag should be generated. Populating the tag is easy task as such most frequent words will be selected as per N-gram algorithm, nevertheless the way they are assigned in Tag cloud is an important task , which uses following important methods or properties that is required for Tag cloud generation.[7]

1. Create a Cloud object and set its properties. One of the most familiar properties is the maximum weight value that defines the range of weight values assigned to tags. It can be set to a convenient value, e.g. the maximum font size. For the minimum weight value can often be kept the default value of zero.

```
Cloud cloud = new Cloud();
```

2. Types of Ordering:

There are four predefined Comparator classes than can be used to specify the way tags are sorted: NameComparatorAsc, NameComparatorDesc, ScoreComparatorAsc, ScoreComparatordesc. You can specify a type of ordering passing an instance of one of these classes to the tags method. For example, to order the tags in descending order of score just call the tags method passing a ScoreComparatorDesc object. By default the tags returned by the tags method are sorted alphabetically, i.e. the NameComparatorAsc is used.

3. The Tag Case:

With the use of setTagCase method you can state how to handle the case of tag names. The possible options are:

- LOWER: tags names are converted to lower case.
- UPPER: tags names are converted to upper case.
- CAPITALIZATION: tags names are capitalized (first letter upper case, other letters lower case).
- PRESERVE_CASE: tags names are not modified and they are case insensitive (e.g. "Home" and "home" are considered the same tag).
- CASE_SENSITIVE: tags names are not modified and they are case sensitive (e.g. "Home" and "home" are considered different tags).

3.3 Twitter Dataset

In this paper, Dataset of Twitter obtainable from SNAP are going to be used that contains 467 million Twitter posts from 20 million users covering a 7 month time from June 1 2009 to December 31 2009. We have a tendency to estimate that this is about 20-30% of all public tweets posted on Twitter during the particular time frame.

For all community tweets, the following information is available: Author, Time, and Content.

Table -1: Twitter Dataset Statistics

Dataset Statistics	
Number of users	17,069,982
Number of tweets	476,553,560
Number of URLs	181,611,080
Number of Hashtags	49,293,684
Number of re-tweets	71,835,017

3.4 Real Tweets Extraction

Using Import.IO's tool named Magic extractor - automatically extracts list of data from a web page using just a URL – there's no setup whatsoever. As per that dataset of real time tweets are been gathered using Magic Extractor. Following is initial page setup where any user's twitter account URL is to be entered in text box. For example –Link of twitter account of Virat Kohli is https://twitter.com/imVkohli that will be parsed through import.io and its tweets are to be extracted.

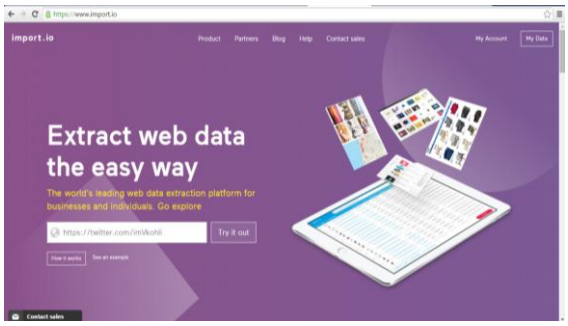


Fig –4: Initial Page of Import.io

Here is snapshot that has extracted Virat Kohli's Twitter tweets and shown in tabular format. It mainly contains its Text header field and its value. Among that for proposed system on portion of Twitter Text and its user name is required thus, that only part of that is to be downloaded. All tweets are downloaded and saved in CSV format.

3.5 Dataset Pre-processing

As mentioned in proposed system some set of pre-Processing required that will create tags which are well thought-out as inputs to tag cloud algorithm. To make sure that, the extracted tags will solitary to consider if it's most frequent and don't seem to be in stop words list of English language, so following procedure is to be pre-processed before producing output of tag cloud.

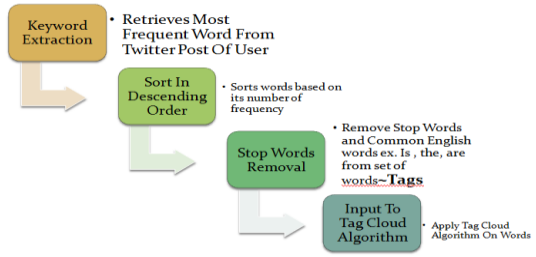
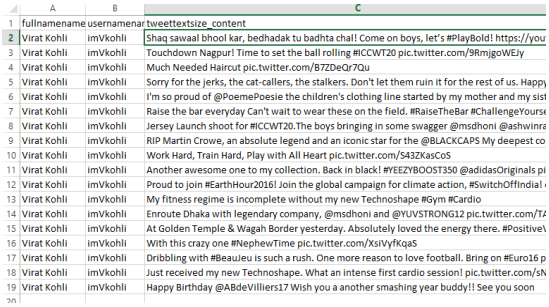


Fig –5: Dataset Pre-processing



	A	B	C
1	fullnamename	usernamenar	tweettextsize_content
2	Virat Kohli	imVkohli	Shaq sawaal bhool kar, bedhadak tu badhta chal! Come on boys, let's #PlayBold! https://you
3	Virat Kohli	imVkohli	Touchdown Nagpur! Time to set the ball rolling #ICCWT20 pic.twitter.com/9RmjgoWEJy
4	Virat Kohli	imVkohli	Much Needed Haircut pic.twitter.com/B7ZDeQr7Qu
5	Virat Kohli	imVkohli	Sorry for the jerks, the cat-callers, the stalkers. Don't let them ruin it for the rest of us. Happy
6	Virat Kohli	imVkohli	I'm so proud of @PoemePoesie the children's clothing line started by my mother and my sist
7	Virat Kohli	imVkohli	Raise the bar everyday Can't wait to wear these on the field. #RaiseTheBar #ChallengeYourse
8	Virat Kohli	imVkohli	Jersey Launch shoot for #ICCWT20.The boys bringing in some swagger @msdhoni @ashwinra
9	Virat Kohli	imVkohli	RIP Martin Crowe, an absolute legend and an iconic star for the @BLACKCAPS My deepest co
10	Virat Kohli	imVkohli	Work Hard, Train Hard, Play with All Heart pic.twitter.com/S43ZKasCoS
11	Virat Kohli	imVkohli	Another awesome one to my collection. Back in black! #YEEZYBOOST350 @adidasOriginals pi
12	Virat Kohli	imVkohli	Proud to join #EarthHour2016! Join the global campaign for climate action, #SwitchOffIndia!
13	Virat Kohli	imVkohli	My fitness regime is incomplete without my new Technoshape #Gym #Cardio
14	Virat Kohli	imVkohli	Enroute Dhaka with legendary company, @msdhoni and @YUVSTRONG12 pic.twitter.com/T
15	Virat Kohli	imVkohli	At Golden Temple & Wagah Border yesterday. Absolutely loved the energy there. #Positive\
16	Virat Kohli	imVkohli	With this crazy one #NephewTime pic.twitter.com/XsiVyfKqaS
17	Virat Kohli	imVkohli	Dribbling with #BeauJeu is such a rush. One more reason to love football. Bring on #Euro16 p
18	Virat Kohli	imVkohli	Just received my new Technoshape. What an intense first cardio session! pic.twitter.com/sN
19	Virat Kohli	imVkohli	Happy Birthday @ABdeVilliers17 Wish you a another smashing year buddy!! See you soon

Fig –6: Tweets of Virat Kohli

In computing, stop words are words which are filtered out before or after processing of natural language data (text).[3] Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. As per proposed system those kind of stop words are not required for Tag Cloud visualization and thus one text file is prepared as like list of words and that to be put included to parse stop words and that words are not been included for Tags which next will be given as to input to Tag cloud algorithm.

3.6 Tag Cloud Generation Algorithm

A tag cloud requires a tag and a number associated with that tag. That number is generally a metric. But, what a tag cloud offers is the ability to get the ordering of both the entity and the metric in a single visual representation. This is done by laying out the data in the order of the entity but changing the size/color intensity of that entity based on the metric value. [9]

As a result, while the user can scan top to bottom (and left to right) for alphabetical ordering of the entities, user can also scan for the font-size/color intensity at the same time. Thus, an extra sort is avoided to get the ordering for each. Of course, for accurate details, one has to sort either for the entity or the metric explicitly.

Next Parameter for generation of tag cloud is of Size of tag/word. This to a large extent depends on the data distribution.

However, sometimes it may not be important showing every entity in the word cloud. Just the top N entities are good enough. So, if we go with the top N approach, then max and the min of the top N entities may not be that wide spread and in this case a linear interpolation should suffice.

As word placement can be somewhat slow for more than a few hundred words, the layout algorithm can be run asynchronously, with a configurable time step size. This makes it possible to animate words as they are placed without stuttering. It is not compulsory to always use a time step even without animations as it prevents the browser's event loop from blocking while placing the words.[10]

The layout algorithm itself is very simple. For each word, starting with the most "important":

1. Attempt to place the word at some starting point: usually near the middle, or somewhere on a central horizontal line.
2. If the word intersects with any previously-placed words, move it one step along an increasing spiral. Repeat until no intersections are found.

Algorithm:

1. Input Extracted Tag K.

2. While T != NIL
   var multiplier = (maxPercent-minPercent)/(max-min);
   var size = minPercent + ((max-(max-(count-min)))*multiplier);

3. Generate Size of Tag
   The least occuring tag(s) will have a font-size of minPercent.
   The most occuring tag(s) will have a font-size of maxPercent.
   Tags with occurence counts in the middle will be scaled linearly.

maxPercent: The font size is set as a percentage. This is the font-size percentage that the largest (most frequent) tag should be set to.

minPercent: This is the font-size percentage that the smallest (least frequent) tag should be set to.

max: This is the number of occurrences for the most frequent tag.

min: This is the number of occurrences for the least frequent tag.

count: This variable should be set inside of the link iterator. It refers to the number of occurrences for the current tag.

Fig –7: Tag Cloud Generated of "imVKohli"

Fig –8: Cover Picture Suggested for "imVKohli" Profile

4. RESULTS & ANALYSIS

There are various online tools that generated tag cloud using online simulation. Although they are good enough but due to lack of N-gram algorithm and stop words are not been removed they will generate tag cloud but more number of words are displayed which degrades the visualization of tag cloud.

Here comparison shown with online tag cloud generator tool -WordItOut lets you transform your text into word clouds, and proposed system implemented tool.

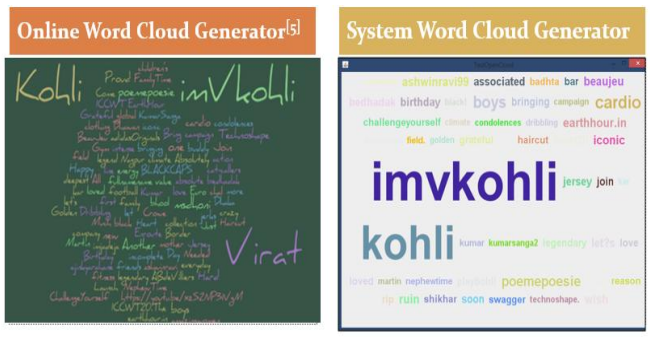


Fig -9: Comparison with online word cloud generator and system word cloud generator

5. CONCLUSIONS

In this paper, we make an innovative use of tag cloud generation algorithm for suggestion of cover picture for twitter social media. Also It has been confirmed by this investigation that it is possible to collect, pre-process, analyze and visualize twitter social data and represent them into more predictable and easy to understandable image format in the form of tag cloud. This will help to identify user personality and save the precious time. This also targets to build tag cloud of user driven module for social community dataset.

REFERENCES

[1] Long Jin, San Diego Yang Chen, Tianyi Wang and Pan Hui, "Understanding User Behavior in Online Social Networks: A Survey," IEEE 2013

[2] C. Wilson et al., "User Interactions in Social Networks and Their Implications," Proc. EuroSys, 2009.Athanasios V. Vasilakos, Kuwait University

[3] "Stopword", https://en.wikipedia.org/wiki/Stop_words

[4] "Import.IO", https://en.wikipedia.org/wiki/import_io

[5] "Web Mining", http://www.web-datamining.net/content/

[6] "Web Content Mining", http://www.cs.uic.edu/~liub/WebContentMining.html

[7] "Open Cloud", http://softron.in/22/53/OpenCloud+-+Getting+Started+Guide.html

[8] "Java", https://simple.wikipedia.org/wiki/Java_(programming_language)

[9] "Twitter", http://en.wikipedia.com/twitter

[10] Clark, J.: Clustered Word Clouds (2008), http://www.neoformix.com/2008/ClusteredWordClouds.html (retrieved on 2009-01-30)

[11] Golder, S., Huberman, B.A., "Usage Patterns of Collaborative Tagging Systems," Journal of Information Science 32, 2, 198–208 (2006)

[12] "NetBeans", https://en.wikipedia.org/wiki/NetBeans