TweetSum: Automated News Summarization of Twitter Trends

Chaitali Khandekar
K.J. Somaiya College of Engineering, India

Raj Daiya
K.J. Somaiya College of Engineering, India

Rhea Parekh
K.J. Somaiya College of Engineering, India

Kavita Kelkar
K.J. Somaiya College of Engineering, India

ABSTRACT
In the recent times, content generated on social blogging sites has become an encyclopedic source of information for every concerned topic in the world[4]. These sites have enabled people to contribute to the vastness of the content available on the internet. During such times, the amount of disorganized and repetitive information generated through these platforms has complicated the means to get the key information. The application that we propose takes an input phrase from the user, captures all tweets related to it and uses them to create a summary from the content already available on the internet.

General Terms
Summarization

Keywords
Summarization, Machine Learning, Social media trends.

1. INTRODUCTION
India being a democratic nation gives the power to the citizen to choose their body of governance. However what is missing right now is platform to know the candidates of the elections. Many a times, we stay anonymous to the portfolio of the candidate standing for the elections. Every right choice even at the root level (e.g. civic body elections) results in the progress of our country and every wrong choice is an obstacle in this progress. Hence access to information about the candidates leads the citizen to an informed choice.

With the growth of internet, "access" to information is not an issue however the availability of resources from different entities at one place is what we propose in this system [3]. Our system will provide the gateway to access all the information available on different cloud storages in the form of an interface which is citizen-friendly. Our system characterizes important aspects of cloud computing and details out the user interface and its interaction with the servers and the data storage systems of cloud.

2. EXISTING SYSTEM
A massive amount of content, related to any subject, is present in form of various blogs, websites, social media posts, news articles etc. on the World Wide Web. The information content multiplies with every passing second, due to which it becomes increasingly inconvenient to find pertinent information. Summary of this information content will assist the user to get an overview of the content in a short span of time. This summarization entails to retrieval of information and generating a condensed interpretation of the retrieved information.

Much research is present on the various techniques of summarization. The system we propose summarizes on the basis of highest frequency key words used in tweets posted relevant to the topic searched and captures pre-existing news articles, Wikipedia pages etc. related to the subject and summarizes that content. Our system, in brief, involves extraction of topic relevant tweets and content present on available websites, followed by generating highest frequency keywords from the extracted tweets and using them to summarize the extracted content.

Our system enables the summary to capture the user interests as it captures and weighs the values of every subtopic related to the subject through its popularity on twitter and only uses the most prevalent subtopics in summary.

3. EXISTING SYSTEM
Most of the summarization systems available today use Natural Language Processing techniques to generate the summary [3]. However, NLP has its own set of limitations. Firstly, NLP has been used effectively only on single or multiple document summarization. Due to which the amount of content processed is limited and poorly extensible to the online content summary system that we are proposing. Secondly, NLP is often combined with Bayesian models to generate contextual sentences. The magnanimous amount of processing that this pertains to is complicated and time consuming with unsatisfactory results [1].

The summary generated by these systems does not show the relationship between the viewpoint of the users and the content available, which is the main aim of our system. The summary generated in these systems does not capture what the users talk about or want to know about. It summarizes only on the basis of the content available in the document.

The system that we propose uses a wide range of information from various sources for this summarization and does the actual summarization based on the ideas of the users[2][3]. These ideas are captured using the user tweets on the micro-blogging website Twitter and the content used is extracted from available contextual content.

4. ARCHITECTURE
Our system consists of two main components that deal with Twitter repository and web scraped content cloud to extract the required contents from their respective databases by the means of HTTP based APIs.

4.1 Twitter API
Access to the real time Twitter feed is required for the summarization of the trending tweets. By the means of third party application such as REST API, users can access the web interface and can extract information or perform their required task. REST API assists any application by providing access to read and write Twitter content. The application via REST API can create a new tweet, extract posted tweets, read public user profiles, obtain follower data etc.

Our proposed system has used OAuth to access the REST API. One of the main reason for using OAuth is that it is application-only authentication; the application makes API

request on the behalf of the users therefore the users are not required to share their login credentials with a third party application thereby maintaining user security. Our system restricts the API to retrieve tweets on a trending topic. The number of tweets retrieved with the help of this API is around 450 posts in a time span of 10 seconds.

These retrieved twitter posts are stored in text format.

4.2 Web Scraper

This component of the system deals with extraction of all the content related to the entered search term from the web. This can broadly include Wikipedia pages, news articles, blog posts, etc. To capture a sense of all the content available related to the entered search term, the process is carried out in two levels.

At the first level is a typical 'Google Search' through the programming interface. For e.g. R programming makes use of the 'getGoogleURL' and 'getGoogleLinks' functions to search for the given topic. It mimics a normal web search but directly posts the search results on the R interface. These results typically include a Wikipedia page followed by popular and recent write-ups on the topic. These results form the basis of the summary.

At the second level of processing two to three links are selected from the Google search results and the content on these web pages is scraped for summary. The RCurl package in R programming is one of the already available packages that allows content extraction on supplying it with a URL. Hence, at this level of processing, we get all the web content available on the World Wide Web from trusted sources and pertaining to the topic.

4.3 User

The primary stakeholder of the architecture is the user. The user enters the phrase for obtaining summary and hence is the primary source for the input. Further, phrase specific summary, top tweets, data frequency cloud etc. are returned to the user.

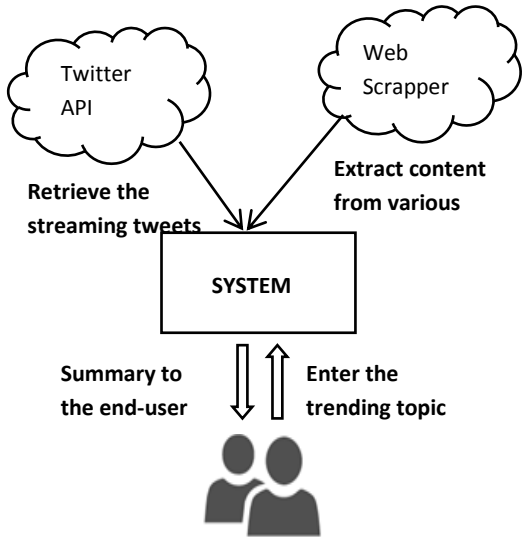

Figure 1: Architecture

5. WORKING OF SYSTEM

Our proposed system captures the context of the topic by summarizing pre-available Web content on the basis of features generated from user tweets. It hence nullifies the need for natural language processing [3]. The entire process entails to three major steps: Extraction, Filtration and Summarization as detailed herewith.

5.1 Extraction of Tweets from Twitter

This is the first step of the system. The user enters the topic for summary on the interface which is then carried forward for extraction from twitter.

Publicly available Twitter third party applications are used for this extraction of tweets. OAuth is the official API used for extraction. The API is supplied with the string variable that stores the phrase and over 450 posts are retrieved in a span of 10 seconds. The API performs a sequential search for tweets based on the phrase and hence 450 tweets containing the phrase are returned. The tweets are initially stored in a .json file. These tweets are then used by loading into a text file for further processing.

5.2 Extraction of Web content

The distinguishing feature of our proposed system is the usage of already available content to form the summary. The available content that the system uses is information available on the Internet related to that topic. The pages are stored in the form of a text file and used later during the summarization step.

5.3 Filtration

The tweets collected after extraction are then processed further to remove non-English tweets, spams, rants, irrelevant data and other sources of misinformation. The system removes all the non-English tweets, removes stopping words, stems those tweets, associates a level of significance on the tweets by influential users, and generates a frequency count of all the words after filtration; thereby generating a set of highest repeating phrases which have the maximum weight from the collected tweets[3].

5.3.1 Stop word elimination: The process of removing certain words unnecessary to text processing is the process of stop word removal. Examples of stop words are "want", "who", "the" etc.

5.3.2 Stemming: Stemming is the process of removing suffixes from words to get the common origin. It basically reduces words to their root form. For example, a stemming process reduces the words "moving", "moved" and "movement" to the root word, "move".

5.3.3 Removal of non-English tweets: The collected tweets are restricted to English language only for further processing. All the tweets in any language but English are eliminated during this process.

5.3.4 Removal of Twitter characters: This process rejects all the special characters (@, #) present in those collected tweets which are irrelevant for the summarization process, but occur in almost all the tweets. Hence such characters are removed so that they do not make it to the top features for summarization process.

After processing, these tweets are converted into a corpus consisting of only relevant information from the tweets so as to optimize the summarization process. Later, the corpus consisting of the collected data is converted into a data-frequency matrix which consists of the top features along with the frequency with which they are being repeated in the tweets. These features are stored in descending order with respect to their associated frequency so as to generate the top "n" features that can be carried forward for the summarization

process. The topmost repeated words having the highest frequency form the base for the proposed summarization process. The output after this stage is the processed data frequency matrix.

The significance of this step is to remove the clutter from the tweets available to gain clarity on the focus area corresponding to the subject.

5.4 Summarization

As mentioned in 5.2 the content extracted from the web is stored as a document. After segmentation of that document, each sentence present in the document is represented as an individual document. The top frequency words obtained in 5.3 are then mapped to each document in order to assign weights to the documents. After the weights are assigned to each document, the top 'n' weighing documents are selected for the summary. The sentences present in the selected documents are printed as a summary for the keyword entered.

Example

Wikipedia Document (The example uses Wikipedia as the base for summary):

Donald John Trump is an American businessman, television personality, politician, and the 45th President of USA. Trump won the general election on November 8, 2016, in a surprise victory against Democratic opponent Hillary Clinton. Trump announced his campaign slogan, "Make America Great Again". Trump first publicly expressed interest in running for political office in 1987. In June 2015, Donald Trump launched his campaign for the 2016 presidential election, and quickly emerged as the front-runner among 17 candidates in the Republican primaries.

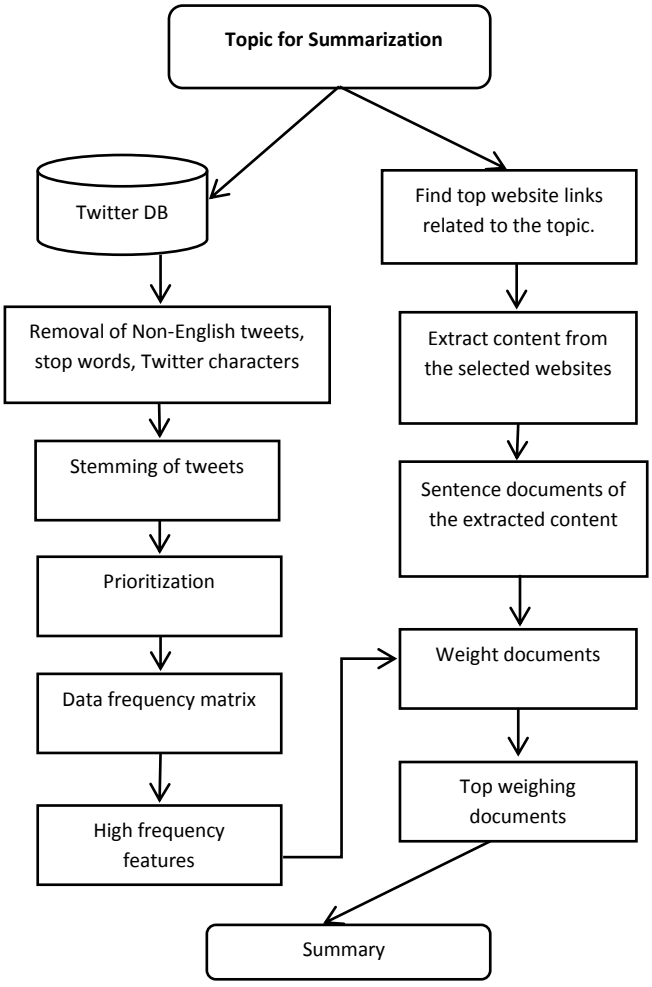

Figure 2: Flowchart

After Segmentation:

Document 1: Donald John Trump is an American businessman, television personality, politician, and the 45th President of United States of America.

Document 2: Trump won the general election on November 8, 2016, in a surprise victory against Democratic opponent Hillary Clinton.

Document 3: Trump announced his campaign slogan, "Make America Great Again".

Document 4: Trump first publicly expressed interest in running for political office in 1987.

Document 5: In June 2015, Donald Trump launched his campaign for the 2016 presidential election, and quickly emerged as the front-runner among 17 candidates in the Republican primaries.

Top frequency keywords:

Keywords	Weight of each keyword
"Trump"	5
"Election"	2
"President"	2
"America"	3

After mapping and assigning weights:

Document	Weight
Document 1	13
Document 2	7
Document 3	8
Document 4	5
Document 5	9

We need a summary of 3 lines (say) , we therefore print the top three weighing documents.

Summary:

Donald John Trump is an American businessman, television personality, politician, and the 45th President of United States of America.

In June 2015, Donald Trump launched his campaign for the 2016 presidential election, and quickly emerged as the front-runner among 17 candidates in the Republican primaries.

Trump announced his campaign slogan, "Make America Great Again".

6. CONCLUSION

In today's world of explosive news and trends sweeping the entire world and the generation of plethora of views and content for every topic, summarization of trends has become the need of the hour. This summarization can be deemed useful only if it provides a contextual summary of the topic after taking into consideration the user point of view.

While many algorithms exist for summarization of single document and even multiple documents[2], content in the document itself is used for summarization. This cannot be extended to the Web content as it ignores the users in the process. If summarization is done giving weightage to what the users consider important then the summary can be deemed useful.

This decision of giving importance to specific content from within the available content is done using user tweets pertaining to that content subject. These tweets enable the weighing of content important to the users and hence make the summary concise and to the point.

By using already available contextual and concise content the system eliminates the need for Natural Language Processing (NLP) hence removing much processing needed to be done on the content. The system leverages the existing editorials and summarizes it based on user's importance.

Going forward Twitter tweets could be replaced by a combination of all social media platforms capable of capturing the user reactions and can provide a more holistic view to the summary generated.

This algorithm aims to fulfill the need of a quick summary of pre-existing web content based on user views and user interactions via social media websites.

7. ACKNOWLEDGMENTS

Our thanks to all the professors of our University for their guidance and their prompt responses to every query we had.

8. REFERENCES

[1] Dunwei Wen, Geoffrey Marshall (2014), *Automatic twitter Topic Summarization* [online], Available: http://ieeexplore.ieee.org/document/7023580/#full-text-section

[2] David Inouye, Jugal K. Kalita (2011), *Comparing Twitter summarization algorithms for Multiple Post Summaries* [online], Available: http://ieeexplore.ieee.org/abstract/document/6113128/

[3] Mr. G. S. Mane, Mrs. A. R. Kulkarni (2015), *Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features* [online], Available: http://www.ijritcc.org/download/1423973129.pdf

[4] Srishti Sharma, Kanika Aggarwal, Palak Papneja, Saheb Singh (2015), *Extraction, summarization and sentiment analysis of trending topics on Twitter*, Available: http://ieeexplore.ieee.org/document/7346696/