News Aggregation Social Bot using Data Mining

Asst. Prof. A. Murugan, Rahul Garg, Aditya Sanjay Ekbote
SRM University

Abstract—The purpose of the system is to showcase the technology developed for building a fully automated real-time news provider. It mainly focuses on using data extraction, data mining and article summarizing algorithm to build the product. The main target of the research is to develop a fully automated platform which can provide news on latest trending topics on social media without wasting much time on different platforms and reading whole article.

Keywords: Text Mining, Web parsing, Text Summarization, News Aggregation

I. INTRODUCTION

From the past decade or so social media has become an important part of almost every ones life. These platforms have evolved from being a mere platform to virtually reach out to the friends to a customised platform for online marketing and getting in touch with the information regarding the surrounding and also around the world. These platforms have a powerful impact on everyone’s life on both personal and professional level, hence data mining algorithms can play a huge part in extracting information from these sources and utilising this information for various applications.

Data extraction has proved to be great area of research in these emerging years. There has always been need of data sources which can be used for extracting the existing data and utilise it for better application. Extracting this data involves parsing through different websites for a specific data with the help of keywords and also the data cannot be used as it is extracted due to the HTML tags and other scripts used inside a web page with data to structure a complete web page hence a fully formulated data extraction algorithm is needed to find that specific vital part of information needed for that application. In this research paper we have also worked with some existing data extraction algorithms and modified them to be work within different parameters such as keywords, date and time and given sources and be fully automated.

Text summarizing deals with extracting the main points of importance from a given source. Initially text summarizing gained importance due to lack of storage media and high cost of storage. But in recent years it has gained popularity due to rapid increase in the quantity and complexity of the data. Summary is defined as getting the main extract of the topic from multiple sources. Text summarizing involves finding an appropriate title, recognising the main keywords, ranking each and every statement according to their position or repetitiveness. In scenarios which involve text summarizing from multiple documents this becomes a more challenging task but is beneficial also as the result tend to be more accurate and specific. In this paper we have tried to modify existing text summarizing algorithms which provide better results for news related summarizing.

Hence this research paper deals with developing a social bot that will leverage the ongoing latest trends from social media platforms, connect with relevant news sources, and perform text extraction and summarizing tasks using algorithms and post the latest news onto multiple social forums. Such an automated system which leverages the latest trends in technology is rarely pre-existing and will prove to leverage the potential of of data mining and social media in a relevant context.

1) We present a comparative study and a realistic approach of extracting the trending topics by studying the different data sets obtained and comparing them using qualitative research.

2) Using a modified approach for data extraction using a tree based approach to get a visualized form of data and then using the tree based algorithms to retrieve the required data through the nodes.


II. ARCHITECTURE OVERVIEW

Automatic news summarizer mainly needs two algorithms one for scraping the web intelligently according to requirements and other a text summarization tool able to efficiently work on multiple documents. Apart from these two major algorithms. The separate modules can come together to build a product that can fetch trends, prioritize them, parse news sources for news, summarize the articles and then display them to the end users.

The main objective of the research paper is to showcase how different sets of algorithms can help us build an automated product. The step-by-step approach is as follows:

1) Extract trending topics from various social media websites in a JSON format and ranking them according to priority.

2) Using the trends to search for news articles from relevant news sources.

3) Utilizing the obtained information from multiple sources and use modified data summarizing algorithm to obtain a content specific and ready to use summary.

4) Using the obtained summarized news snippets to build out a product.

We will discuss about each of these modules below:

A. Latest trends extraction

The objective of this module is utilizing the discrete data present in form of hashtags and topic listing on various online forums and social media sites like reddit, twitter, facebook, google trends for collecting the live top trending topics. The trending topic names could conceivably be demonstrative of the sort of data individuals are tweeting about unless one puruses the pattern content related with it.
For instance, #happyvalentinesday demonstrates that individuals are tweeting about Valentines Day. A pattern named Boone Logan is demonstrative that tweets are about individual named Boone Logan. The approach here involves utilize these different information from various sources which is mainly made available through their developer REST API(Application Programming Interface) and then comparing these different information from various sources for prioritizing these topics. To resolve ambiguity of the fetched news sources, it is critical to prioritize the trends and also filter the similar trends. As an example, if #IPL2017 and #IPLisLive both represent the event IPL. Hence, by using Stemming Text Mining Python libraries, we can find the similarity coefficient and hence filter out these words and only choose #IPL as the preferred news trend.

B. Text Summarization of News Articles

This is the main module which summarizes the information crawled through different news sources using a summarization algorithm based on unique sentence quotient and the NLTK library. The algorithm here is designed in such a manner as to enhance the result efficiency with the increase in the sources of information.

A mathematical approach for text summarizing problem is demonstrated here involving using basic mathematical equation to evaluate each and every statement and rank them suitable for summarizing text or not. The main focus of summarization is to eliminate similar sentences and only have unique distinctive sentences in the news article, so as to provide the maximum context for the news in minimum words.

C. Web Scraping for News Articles

The objective of this module is scraping all the relevant information from the pre-determined news sources and compiling the news articles for a particular search, which can then be passed to the summarization module. This mainly involves scraping through HTML code and getting the required result by reinforced learning involved in understanding the structure of a particular website from which the data is to be crawled. The algorithm used is mainly vision based page segmentation algorithm which involves visualizing the whole HTML document as a DOM tree and then using tree traversal and other tree based algorithms to visit different nodes and utilizing them to retrieve the valuable information.

D. Social Bot for content publishing

This is the final module whose function is posting the summarized news snippets on the product front-end. An Internet Bot, also known as web robot, WWW robot or simply bot, is a software application that runs automated tasks (scripts) over the Internet. We can utilize the bot to sync all the modules and hence at determined intervals publish precise news snippets which ill focus on only live trending topics. We can also extend to social media by posting on a particular profile on a timely basis. A particular timestamp is set after which the bot checks for new information from the database and posts that information using the api made possible through these websites in a presentable manner.

III. WEB CRAWLING ALGORITHM

A. Vision Based Page Segmentation (VIPS)

This technique is used for extracting a semantic form of a web page. The tree based structure thus obtained in hierarchical form and each node in it represents a block of HTML code. Each node is assigned a value known as degree of coherence which represents the value of the block on a visual basis. The VIPS calculation makes full utilization of page format include:

It first focuses all the fitting squares from the html DOM tree, then it tries to find the separators between these removed pieces. Here, separators mean the level or vertical lines in a site page that apparently cross with no squares. Finally, in perspective of these separators, the semantic structure for the site page is assembled. VIPS figuring uses a top-down approach, which is outstandingly effective.

For each and every block degree of coherence is calculated which measures how coherence the content is. DoC has the following properties:

1) The greater the DoC value, the more consistent the content within the block
2) In the hierarchy tree, the DoC of the child is not smaller than that of its parent

The value of degree of coherence is in the range 1 to 10.

B. The VIPS Algorithm

Here, we exhibit the VIPS estimation. On a very basic level, the vision-based substance structure of a page is obtained by joining the DOM structure and the visual prompts. The division technique is spoken to in Figure 2. It has three phases: piece extraction, separator area and substance structure improvement. These three phases with everything taken into account are seen as a round. The figuring is beat down. The site page is immediately distributed into a couple real squares and the different leveled structure of this level is recorded. For each huge upset, a comparative division process is done recursively until we get enough little impedes whose DoC qualities are more conspicuous than pre-portrayed PDoC.
For each round, the DOM tree with its visual information identified with the present piece (page for the first round) is gotten from a web program, as we show in Figure 3. By then, from the root node(s) of the DOM tree (e.g. DOM center point 1 in Figure 3b), the piece extraction process is started to focus deter from the DOM tree in perspective of visual signs. Every DOM (center point 1, 2, 3, 4, 5, 6, 7 as showed up in figure 3b) is checked to judge whether it shapes a single square or not. If not (center point 1, 3, 4 in figure 3b), its youths will be taken care of likewise. We will dole out a DoC motivator to each expelled piece (center 2, 5, 6, 7 in figure 3b) in light of the square’s visual property. Right when all squares of the current round in current page or sub-page are isolated, they are put into a pool. Separators among these pieces are recognized and the largeness of a separator is set in perspective of properties of its neighboring squares. The outline pecking request was inherent light of these separators. Resulting to building the arrangement levels of leadership of the current round, each leaf center point of the substance structure is checked paying little heed to whether it meets the granularity essential. If not, this leaf center will be managed as a sub-page and will be also divided moreover.

IV. ARTICLE SUMMARIZATION ALGORITHM

The goal was to combine articles over a similar news topic from multiple sources and then summarize the whole article corpus into a summary so as to have multiple angles to the same story and less similarity between sentences. It was noticed that if a single article was summarized the probability of similarity of sentences was far greater than when multiple articles from different news websites were appended to form the text corpus to be fed into the summarizer.

The algorithm was driven using a simple concept which is to eliminate the similarity coefficient between sentences. It is noticed that to provide maximum information about a news, the article must consist of sentences with different context. The algorithm uses the NLTK library and self deviated formulae to score each sentence in the text corpus and then filter the sentences over an upper and lower bound limits. The sentences are first filtered to remove any stopwords in them. Stopwords are words which are filtered out before any NLP process such that the sentence contains only lexical words. The NLTK has a corpus of English stopwords. The tokenized sentences without the stopwords are then compared and scored along with all other sentences in the text. The score of the sentence should lie within the bounds. The upper bound of the score denotes that the 2 sentences are almost similar or same. The lower bound will be kept so as to avoid any single word sentences, such as the place of reporting the news.

The formulae for scoring two sentences is:

\[ \text{sum}(\text{compare}(\text{sent}, \text{sent1}))/\text{float}(\text{len}(\text{sent})) \]

where, sents are all the sentences in the text, sent1 is iterated sentence and sent is the current sentence. The comparing score given out while comparing the two sentences measures if the two sentences do not have similar words.

\[ \text{len}(\text{sent})/\text{len}(\text{sent2}) \]

The upper bound is kept at around 0.9 and the lower bound is set to 0.10. The sentences are filtered to lie in bounds and then arranged in the order in which they appear in the text, to maintain the contextual flow of the article.

The algorithm is a enhanced approach to the textRank algorithm, specific to summarization of news articles. TextRank will rank sentences over the attribute of their "closeness" in the text. In the case if a news articles, we need to filter such sentences to increase the context output from minimum sentences. The algorithm can be further enhanced to filter various news article related words, such as the place of occurrence, time of occurrence can be figured out by the algorithm and then such sentences which contain them can be filtered for the final summary.

V. CONCLUSION

In this paper a modified approach for web crawling and text summarization is proposed. The modification is made in such a way that both algorithms complements each other's function and provides a modified desired result. From trending topic extraction to web scraping to article summarization this paper utilizes the useful prospects of each algorithm.
and a system is prepared which can have a wide number of applications. The web crawling produces a web content structure which is quite useful in applications such as web adaptation, information retrieval and information extraction. An approach to web crawling by visualizing it as a top down tree is presented which is scalable and can be adjusted according to the structure of a webpage. Contrasted and conventional DOM based division technique, the plan uses helpful visual signs to acquire a superior segment of a page at the semantic level. It is likewise free of physical acknowledgment and functions admirably notwithstanding when the physical structure is far not quite the same as visual introduction. The calculation is assessed physically on a huge informational index, and furthermore utilized for choosing great development terms in a pseudo-importance criticism prepare in web data recovery, both of which accomplish exceptionally tasteful execution. Article summarization algorithm on the other hand provides a mathematical approach for document summarization which increases the chances of better results and hence proves better than other algorithms like road runner and others. Thus the system proposed in this paper provides users with the exact and correct information obtained and verified from various sources and hence result in creation of a fully-automated news generation bot which can be a huge boost to the media industry, Development of yet another social media platform, extracting the potential of social media footfall to generate quality content for app and exploration of new techniques in the field of data mining

REFERENCES

[8] https://dev.twitter.com/docs
[17] Automatic Hashtag Recommendation for Microblogs using Topic-specific Translation ModelZhuo y e Ding Qi Zhang X uanJing Huang School of Computer Science, Fudan University, 825 Zhangheng Road, Shanghai, P.R. China
[18] Atefeh Farzindar; Diana Inkpen, "Natural Language Processing for Social Media,” inNatural Language Processing for Social Media, 1, Morgan and Claypool, 2015